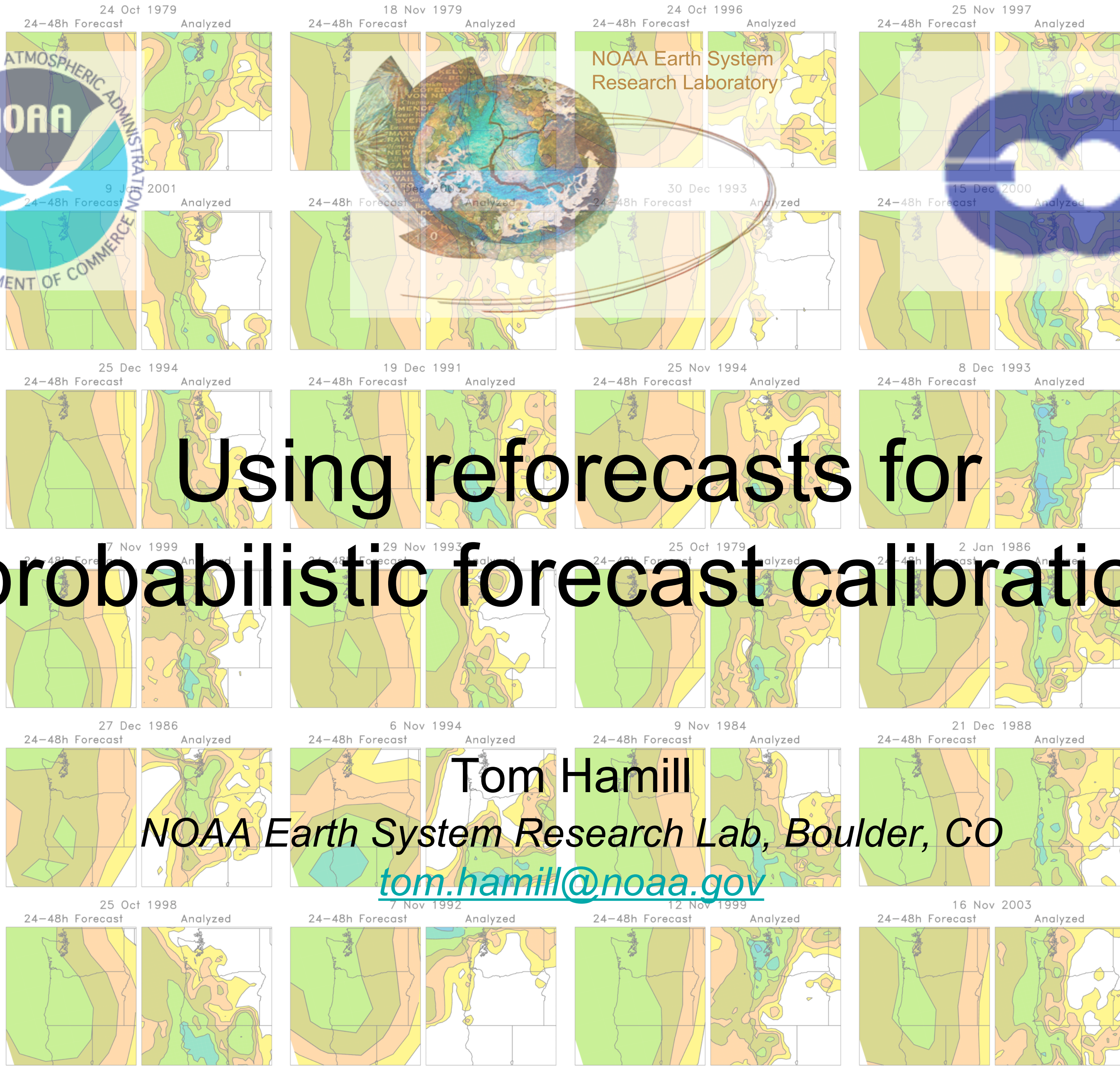


Using reforecasts for probabilistic forecast calibration

Tom Hamill

NOAA Earth System Research Lab, Boulder, CO

tom.hamill@noaa.gov

What is a reforecast?

- A hindcast, a numerical prediction for a date in the past *using the model and data assimilation system that is currently operational*.

Why compute reforecasts?

- For many forecast problems, such as long-lead forecasts or high-precipitation events, a few past forecasts may be insufficient for calibrating the probabilistic forecasts.

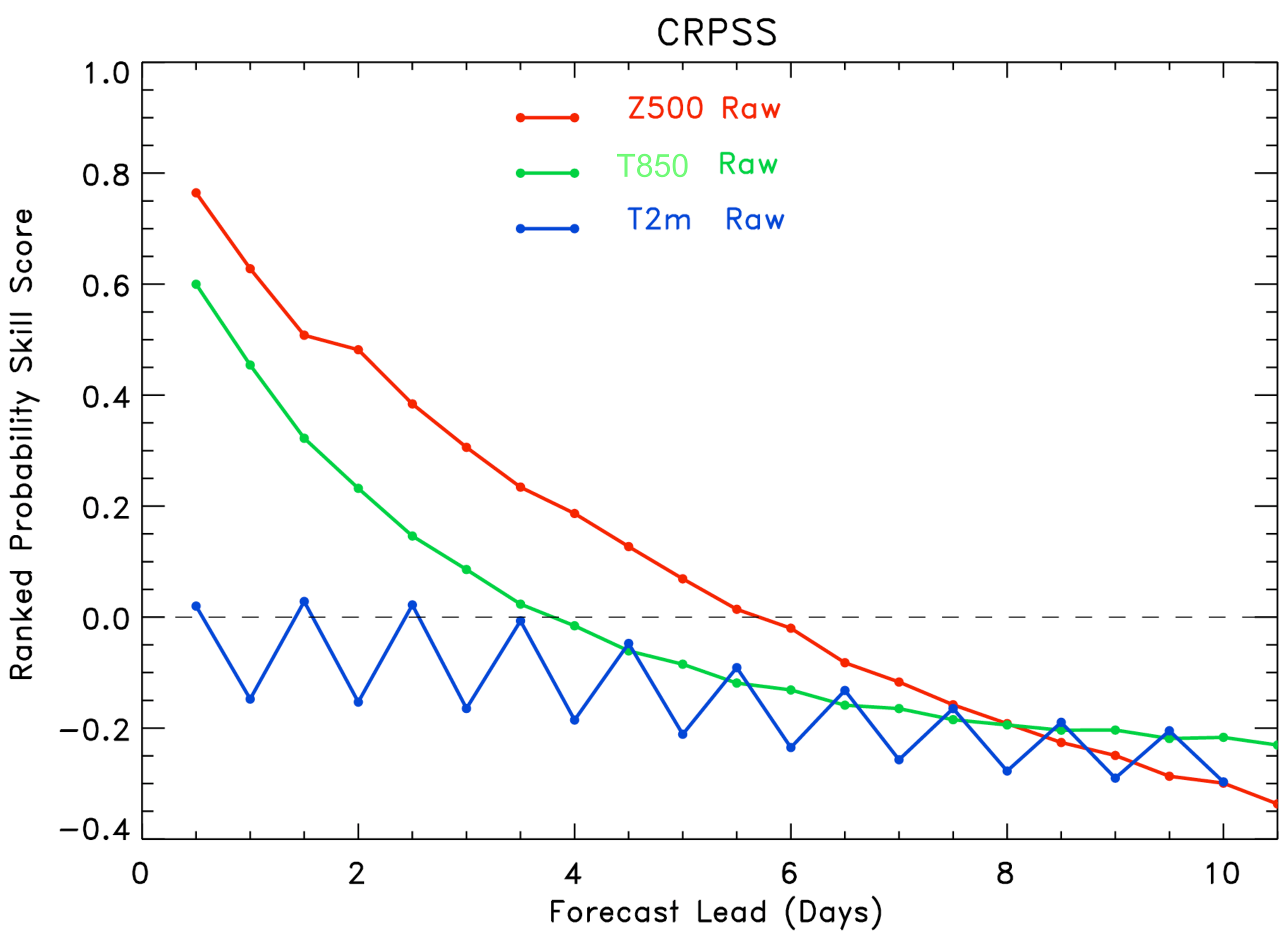
NOAA's reforecast data set

- **Model**: T62L28 NCEP GFS, circa 1998
- **Initial States**: NCEP-NCAR Reanalysis II plus 7 +/- bred modes.
- **Duration**: 15 days runs every day at 00Z from 19781101 to now. (*http://www.cdc.noaa.gov/people/jeffrey.s.whitaker/refcst/week2*).
- **Data**: Selected fields (winds, hgt, temp on 5 press levels, precip, t2m, u10m, v10m, pwat, prmsl, rh700, heating). NCEP/NCAR reanalysis verifying fields included (Web form to download at *http://www.cdc.noaa.gov/reforecast*). Data saved on 2.5-degree grid.
- **Experimental precipitation forecast products**: http://www.cdc.noaa.gov/reforecast/narr .

Outline

- Several applications of 1998 GFS reforecasts.
  - Comparison of Z500, T850, T2m
  - 6-10 day forecasts over US
  - Downscaled PQPF in US
  - Monsoon PQPF in India
  - Tornado forecasts
- An exploration of whether reforecasts from a much-improved 2005 ECMWF model provide similar benefits as were achieved for 1998 GFS

Skill of 500-hPa Z, 850-hPa T, and 2-m T from raw 1998 GFS reforecast ensemble

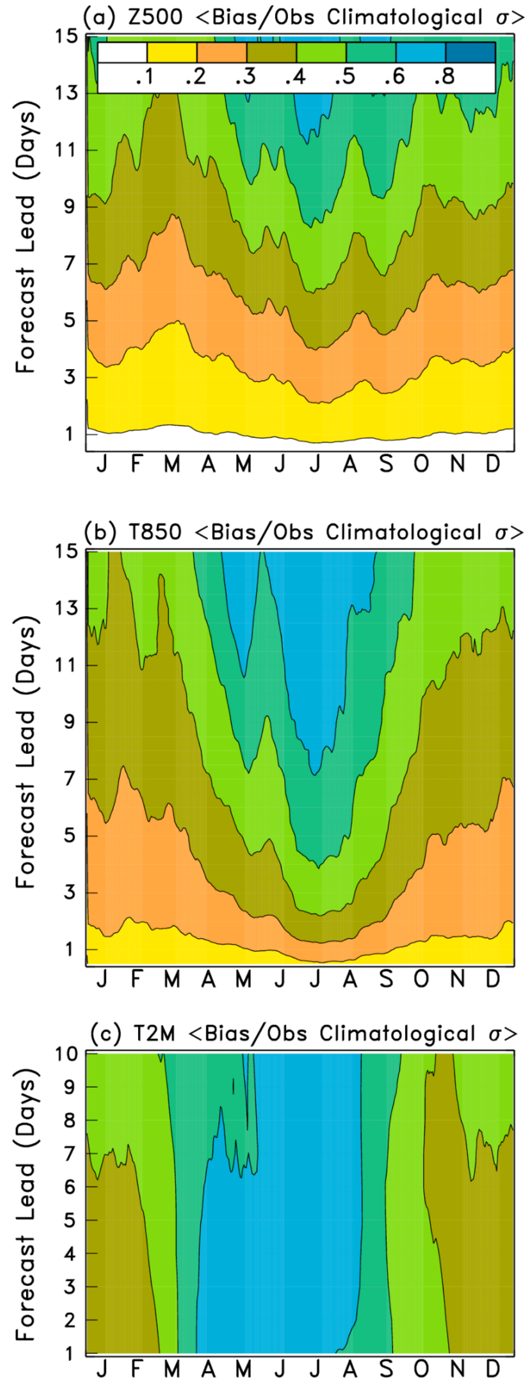


1998 T62 GFS much less accurate than current models, but qualitatively still the same with current models.

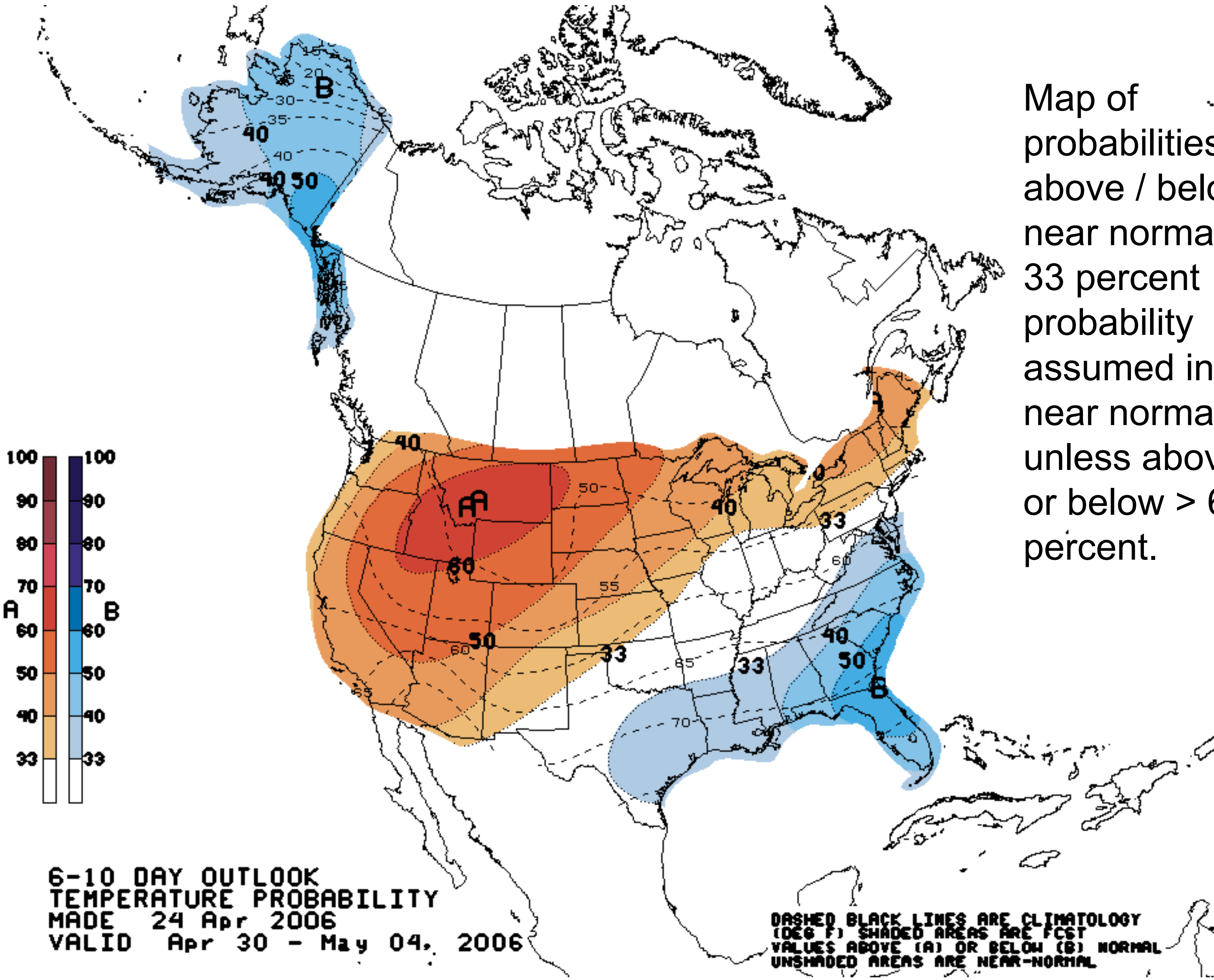
The one we probably care about the most, T_{2m}, scores the worst.

(1979-2004 data)

Forecast bias contaminates T_{2m} much more than Z_{500}

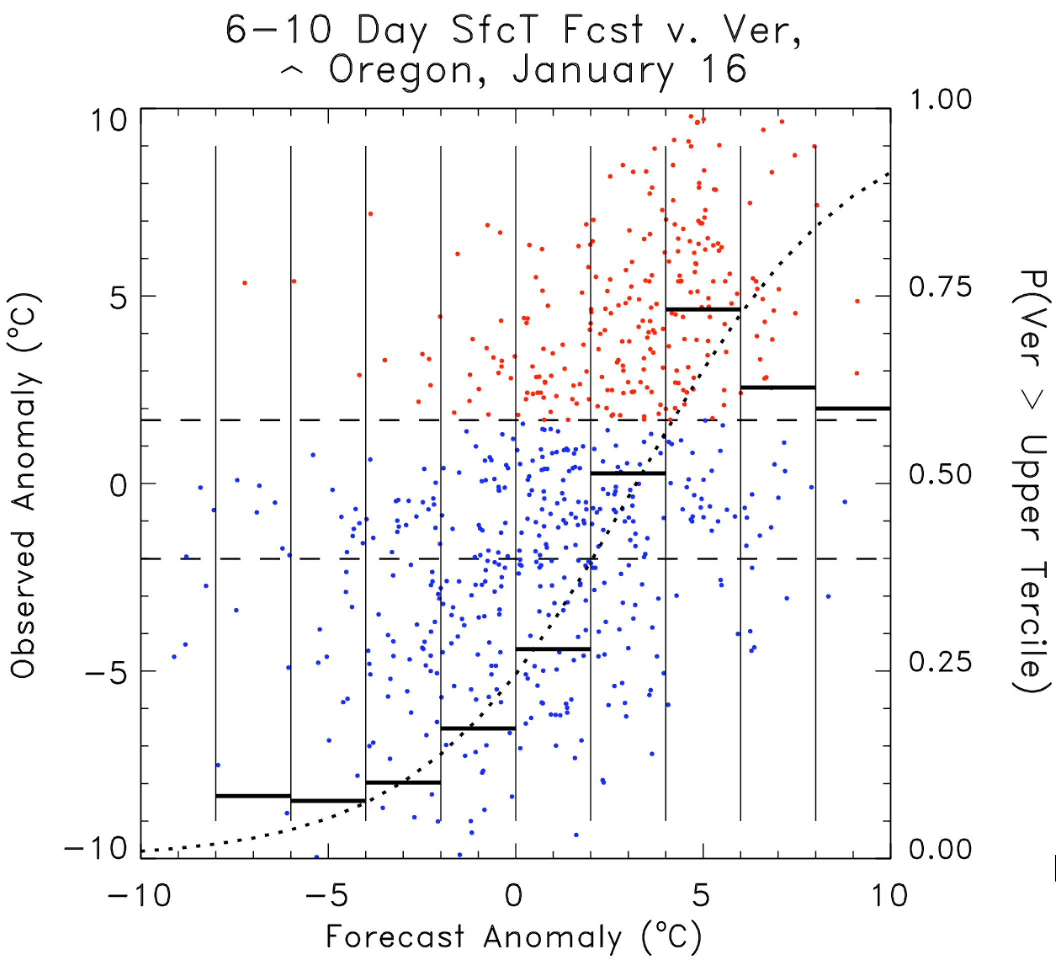


Application: NCEP/CPC's 6-10 day outlook



Map of probabilities of above / below / near normal. 33 percent probability assumed in near normal unless above or below > 67 percent.

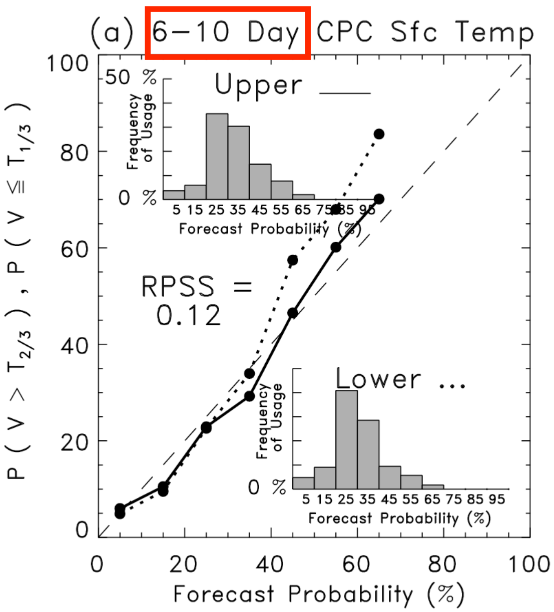
Using a long reforecast data set of observed and forecast anomalies



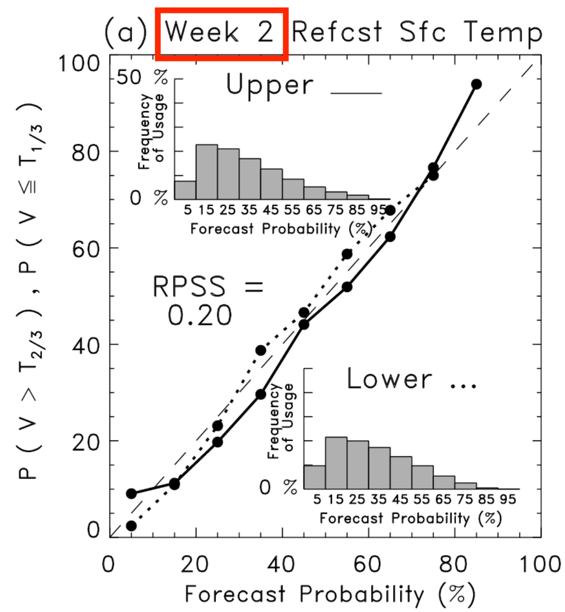
With our reforecasts, we have 25+ years of data. Let's use old data in a 31-day window around the date of interest to make statistical corrections.

Dashed lines: tercile boundaries
Red points: samples above upper tercile
Blue points: samples below upper tercile
Solid bars: probabilities by bin count
Dotted line: a fitted model, TBD

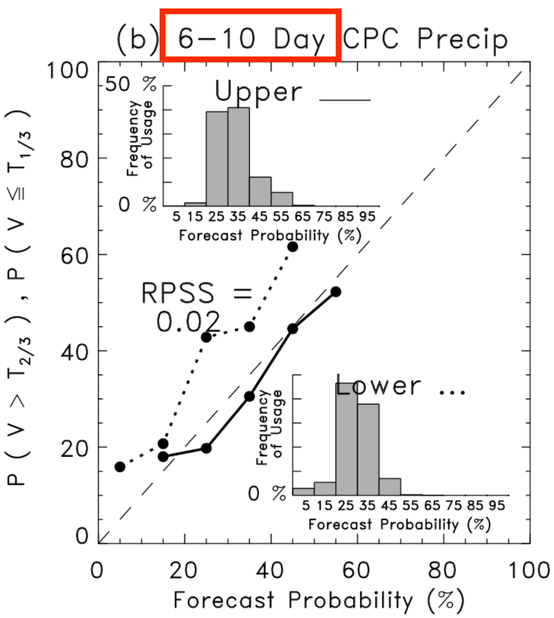
Comparison against NCEP / CPC forecasts at 155 stations, 100 days in winter 2001-2002



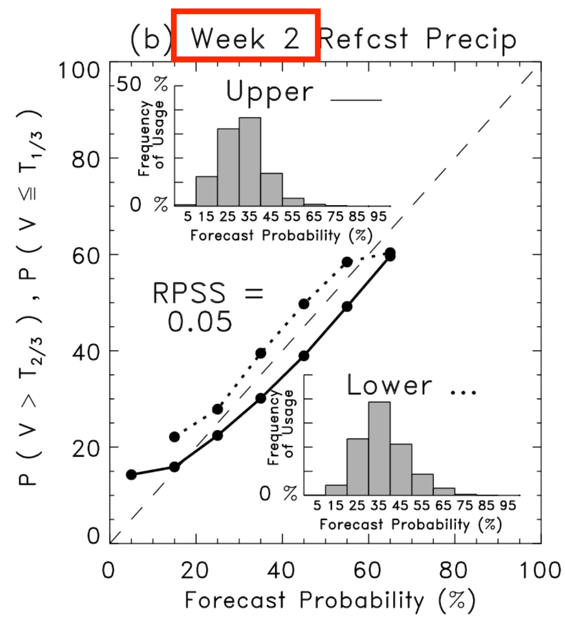
temperature forecasts



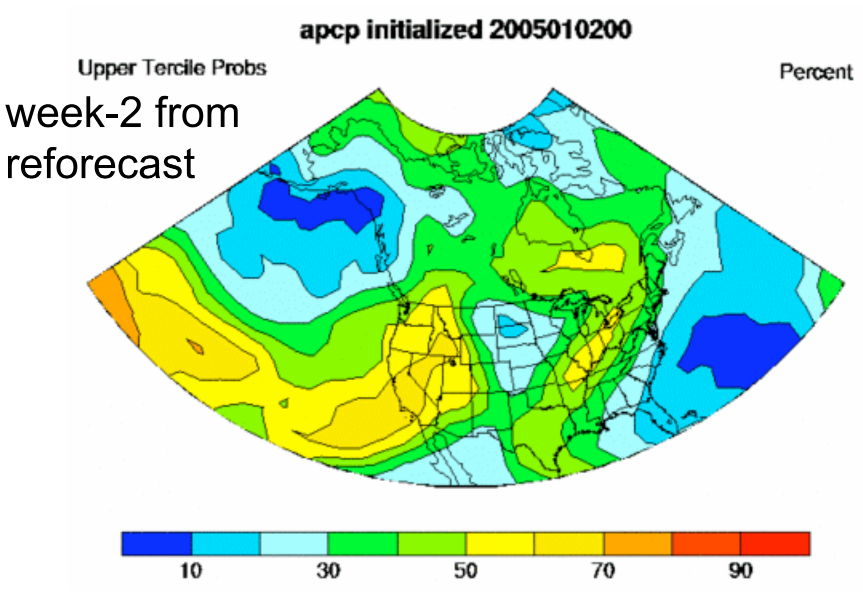
Reforecast calibrated **Week-2** forecasts more skillful than operational NCEP/CPC **6-10 day,** which was based on human blending of NCEP, ECMWF, other tools.

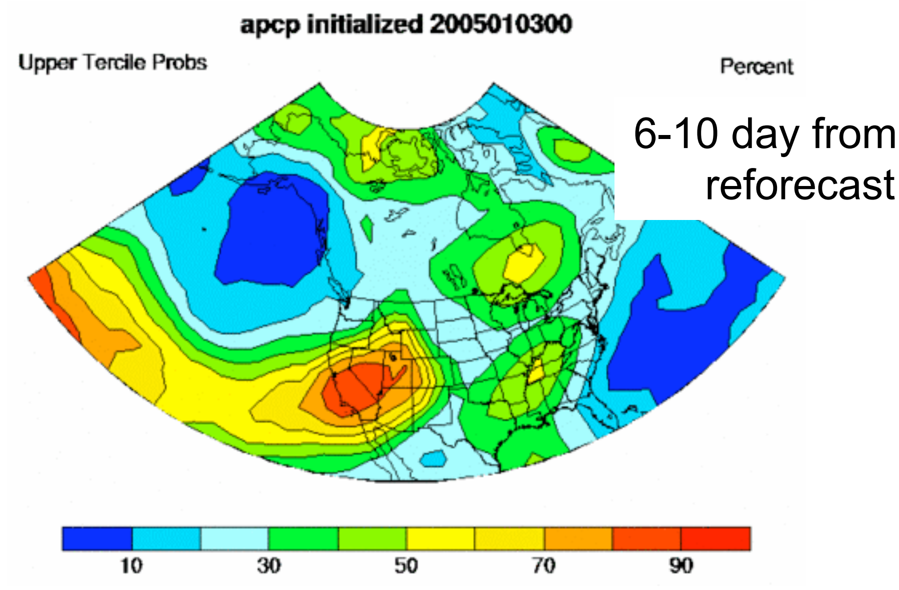


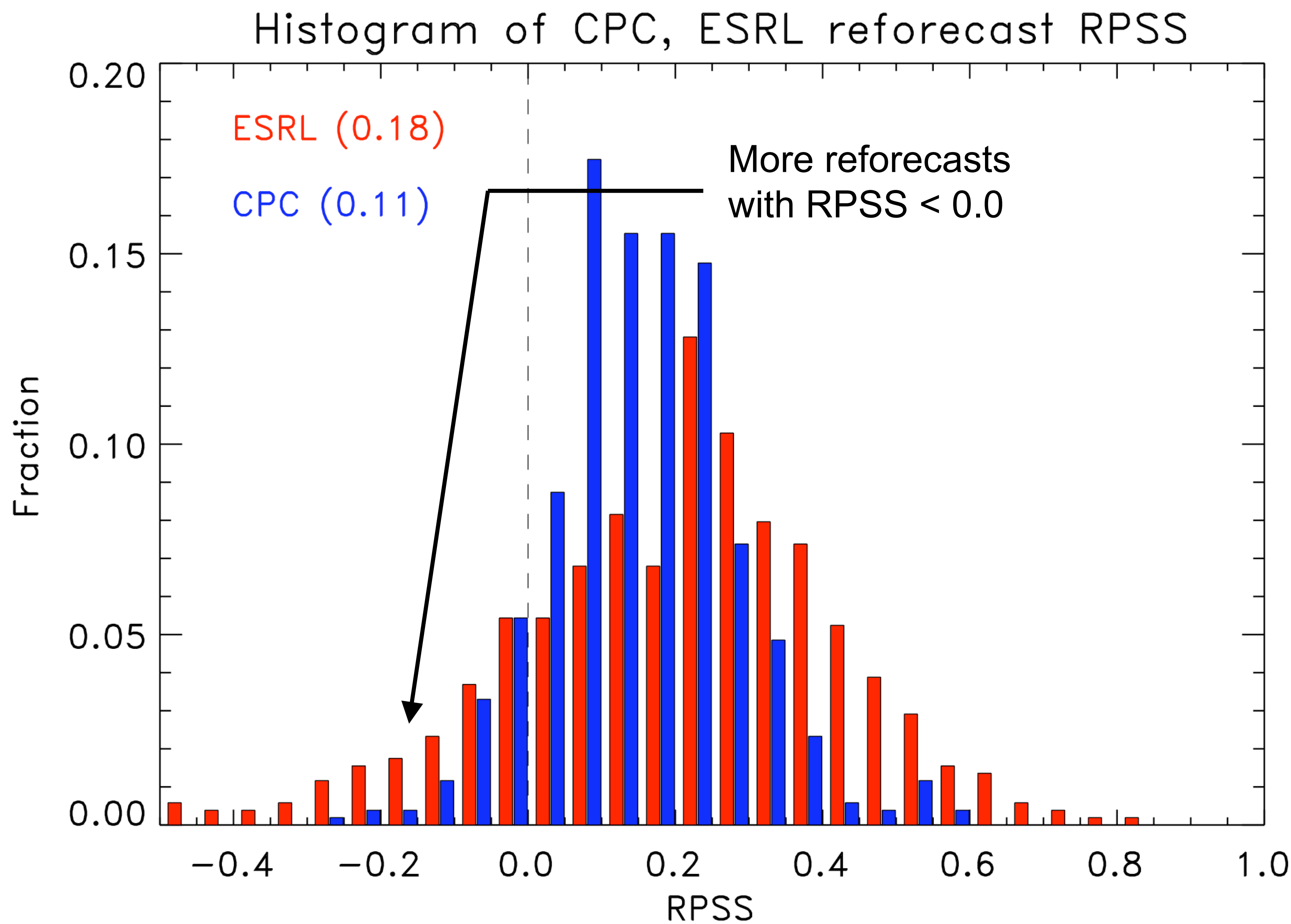
precipitation forecasts

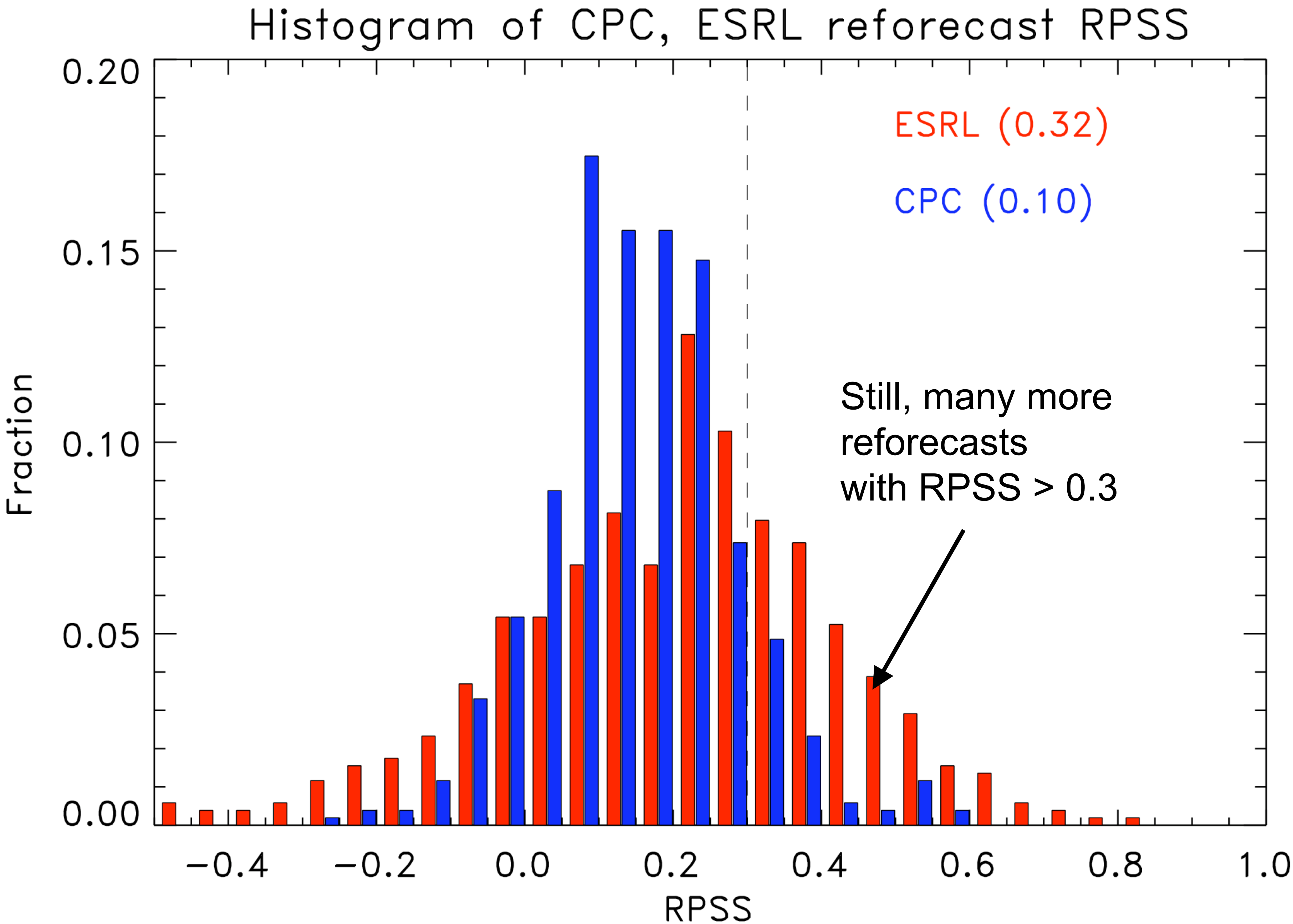


Reforecast-based example: floods causing La Conchita, California landslide, 12 Jan 2005



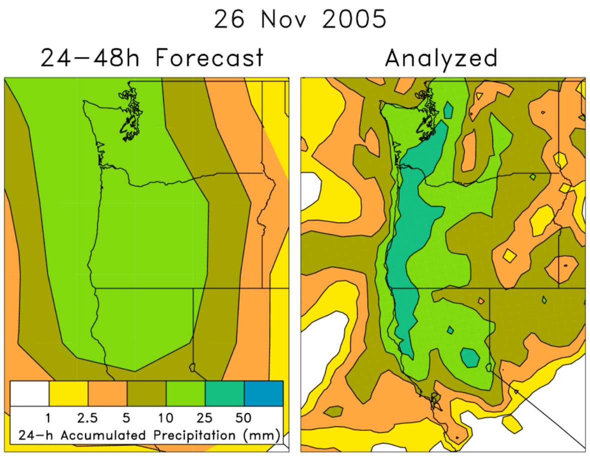


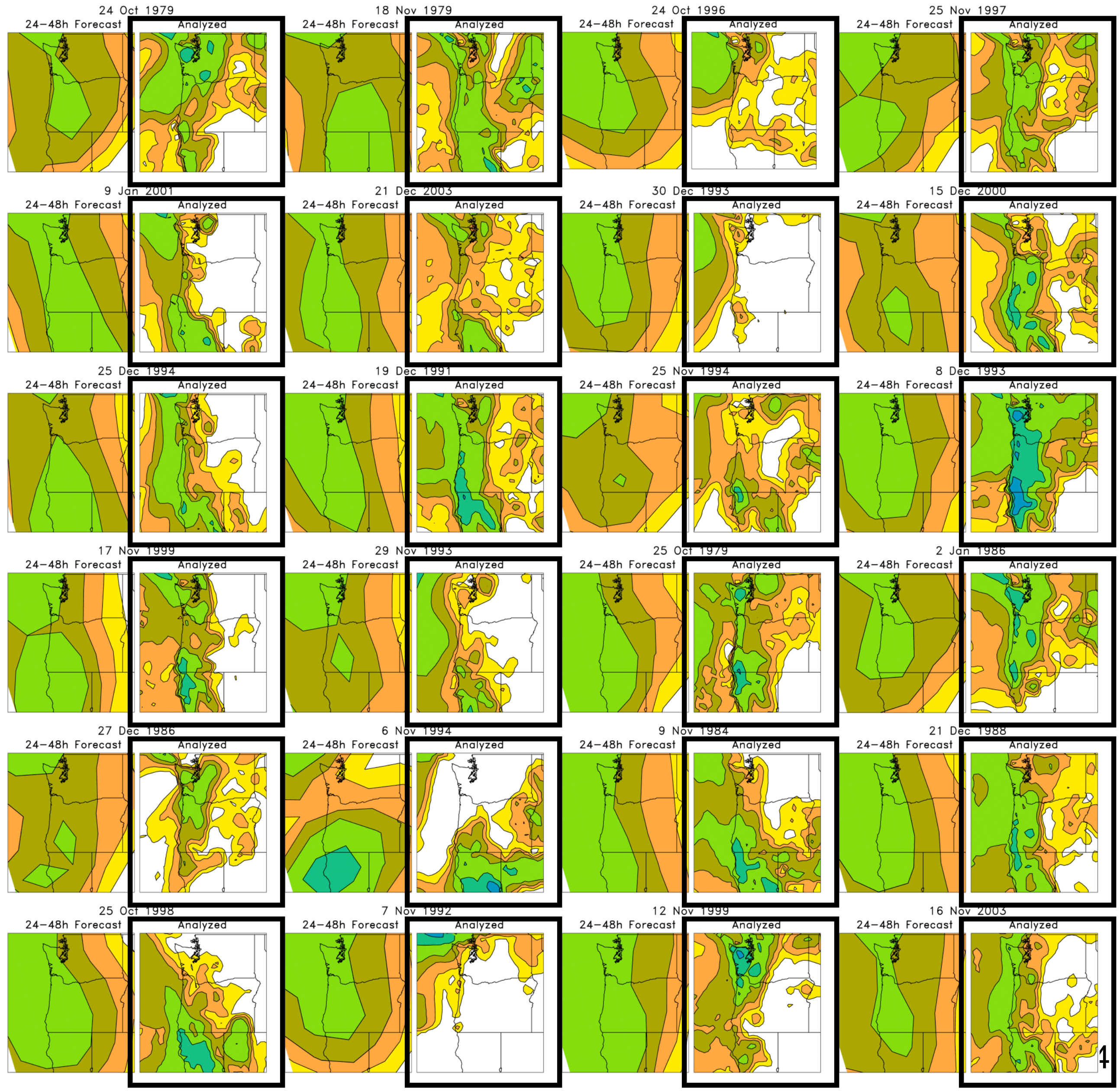

Histogram of CPC, ESRL reforecast RPSS
ESRL (0.18)
CPC (0.11)
More reforecasts
with RPSS < 0.0
Fraction
0.20
0.15
0.10
0.05
0.00
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
RPSS



Histogram of CPC, ESRL reforecast RPSS
ESRL (0.32)
CPC (0.10)
Still, many more
reforecasts
with RPSS > 0.3
Fraction
0.20
0.15
0.10
0.05
0.00
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
RPSS

Application: downscaled precipitation forecasts using analog technique

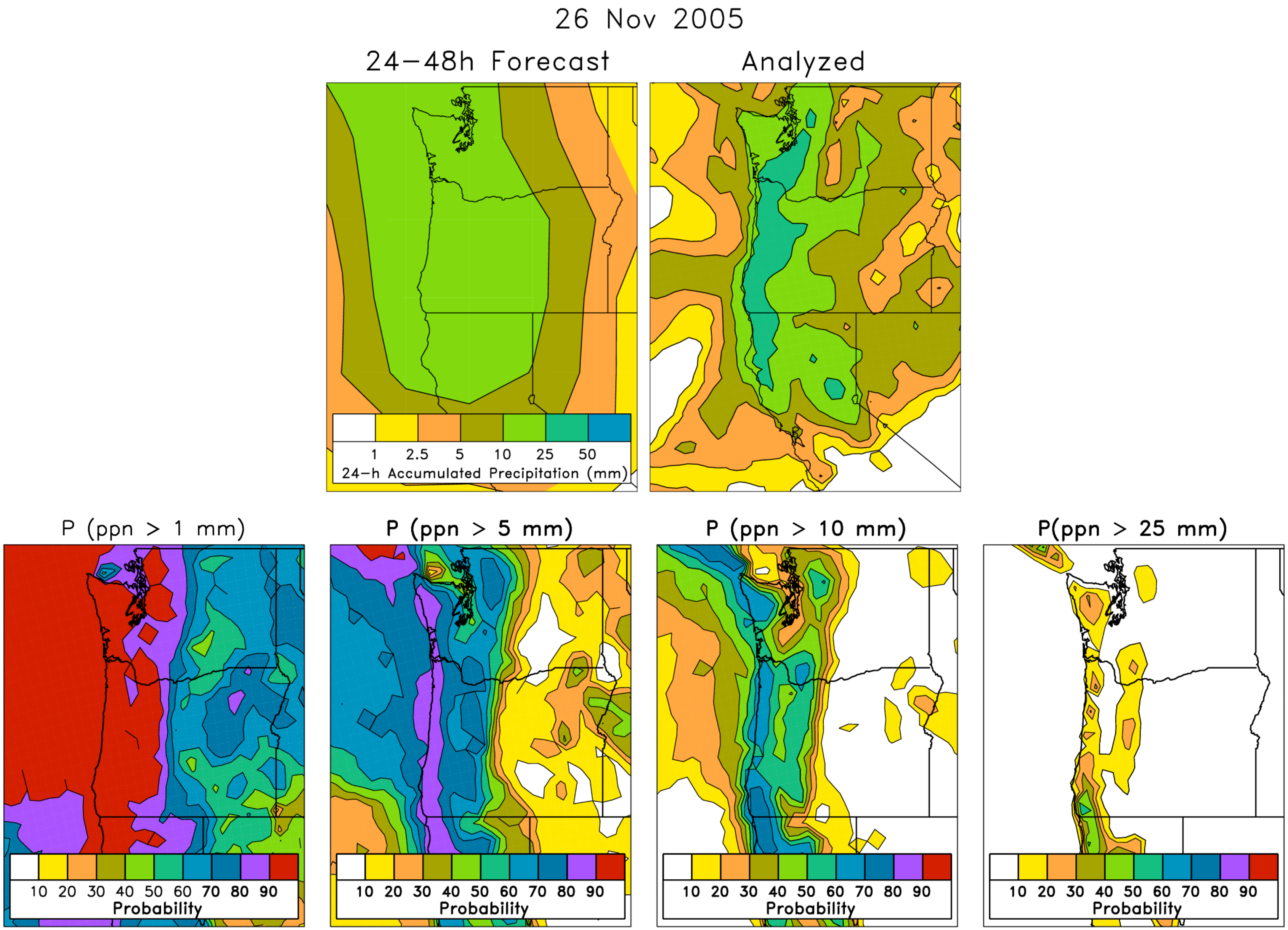


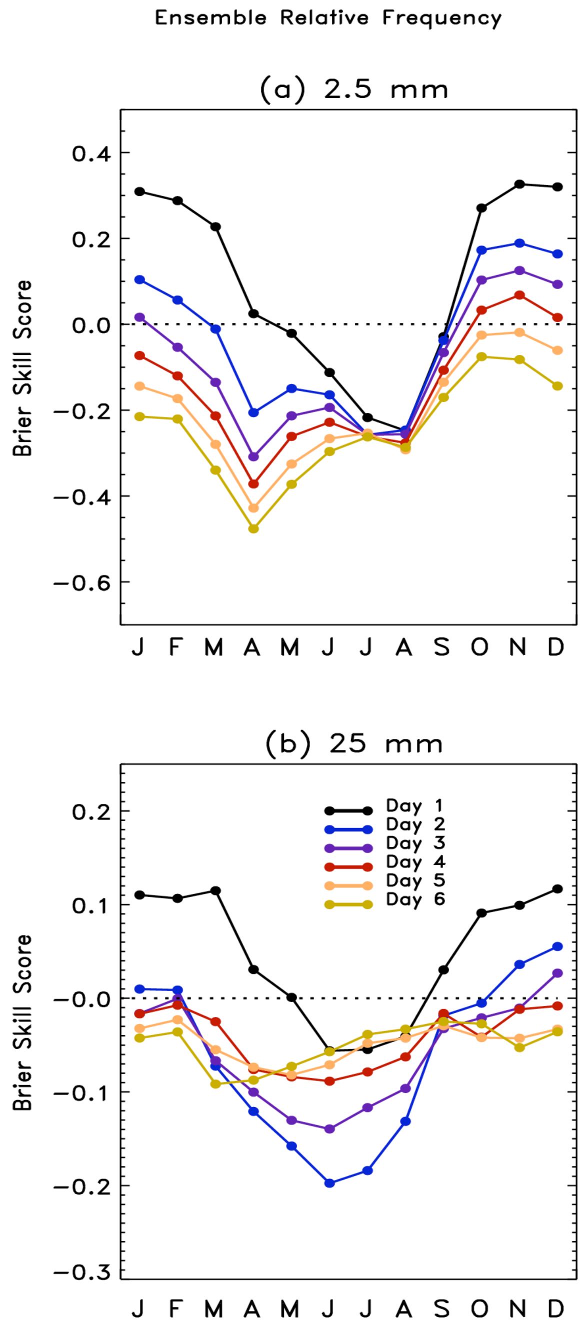


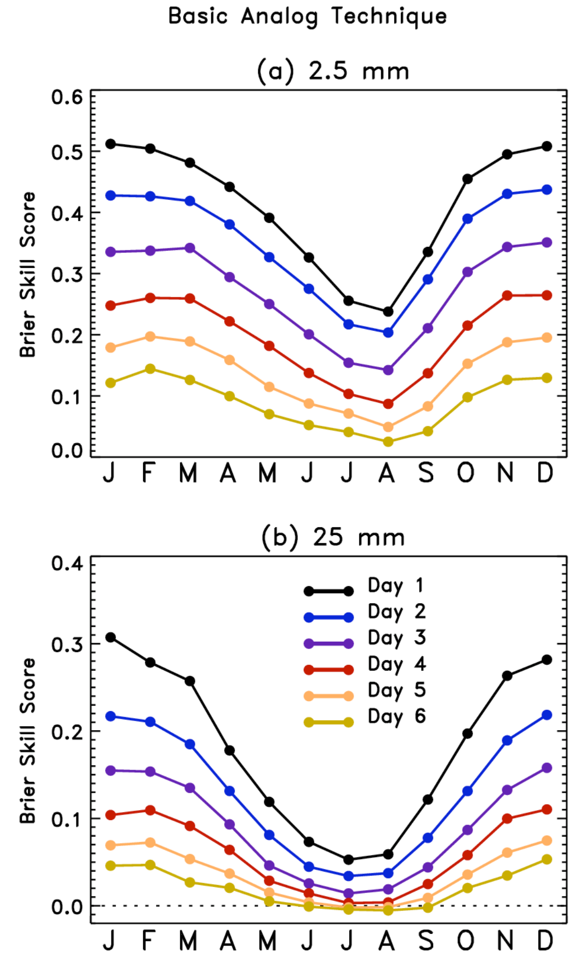
On the left are old forecasts similar to today's ensemble-mean forecast. The data on the right, the analyzed precipitation conditional upon the forecast, can be used to statistically adjust and downscale the forecast.

Analog approaches like this may be particularly useful for hydrologic ensemble applications, where an ensemble of realizations is needed.

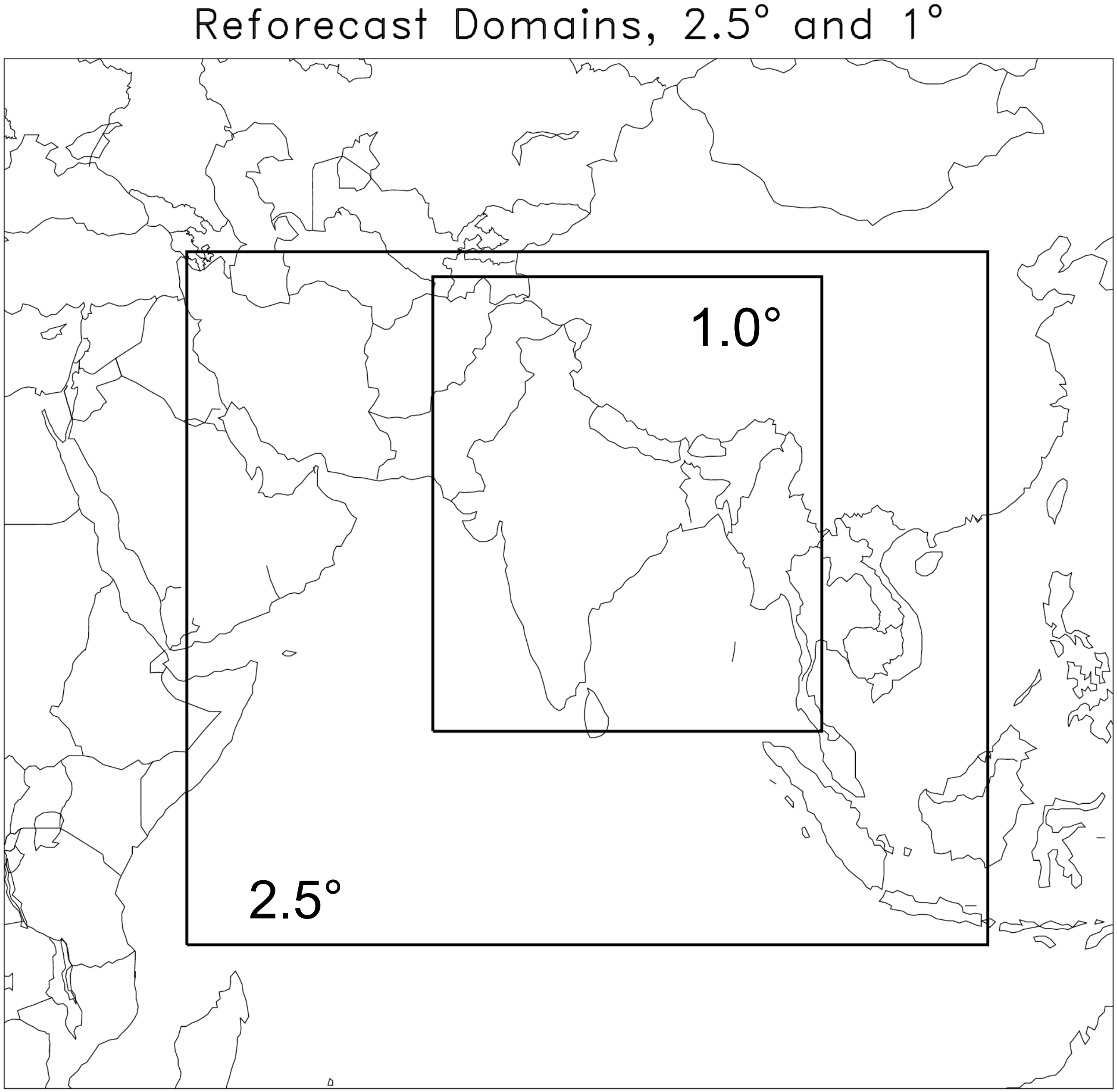
Downscaled analog probability forecasts







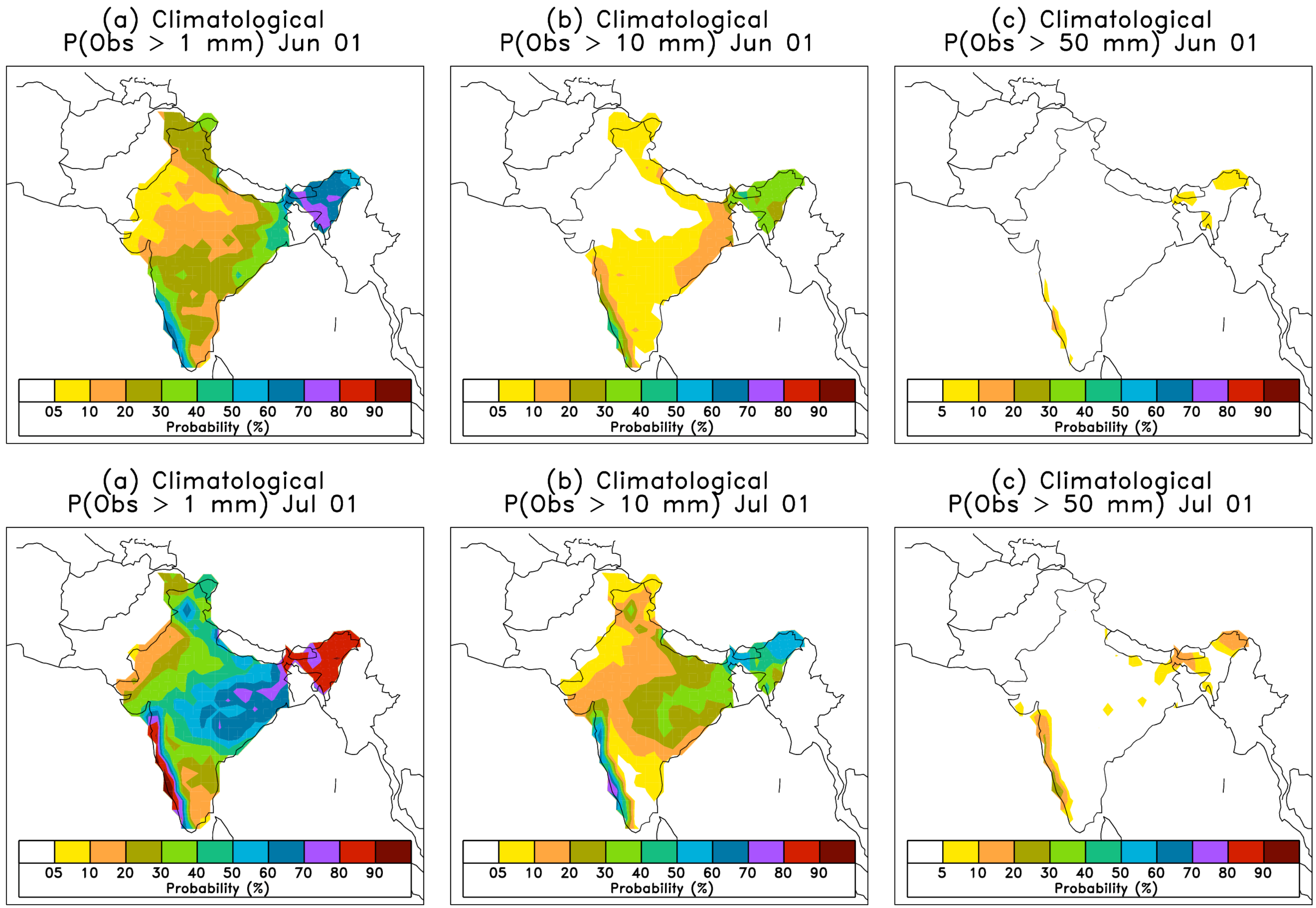
Verified over 25 years of forecasts; skill scores use conventional method of calculation which may overestimate skill (Hamill and Juras 2006).



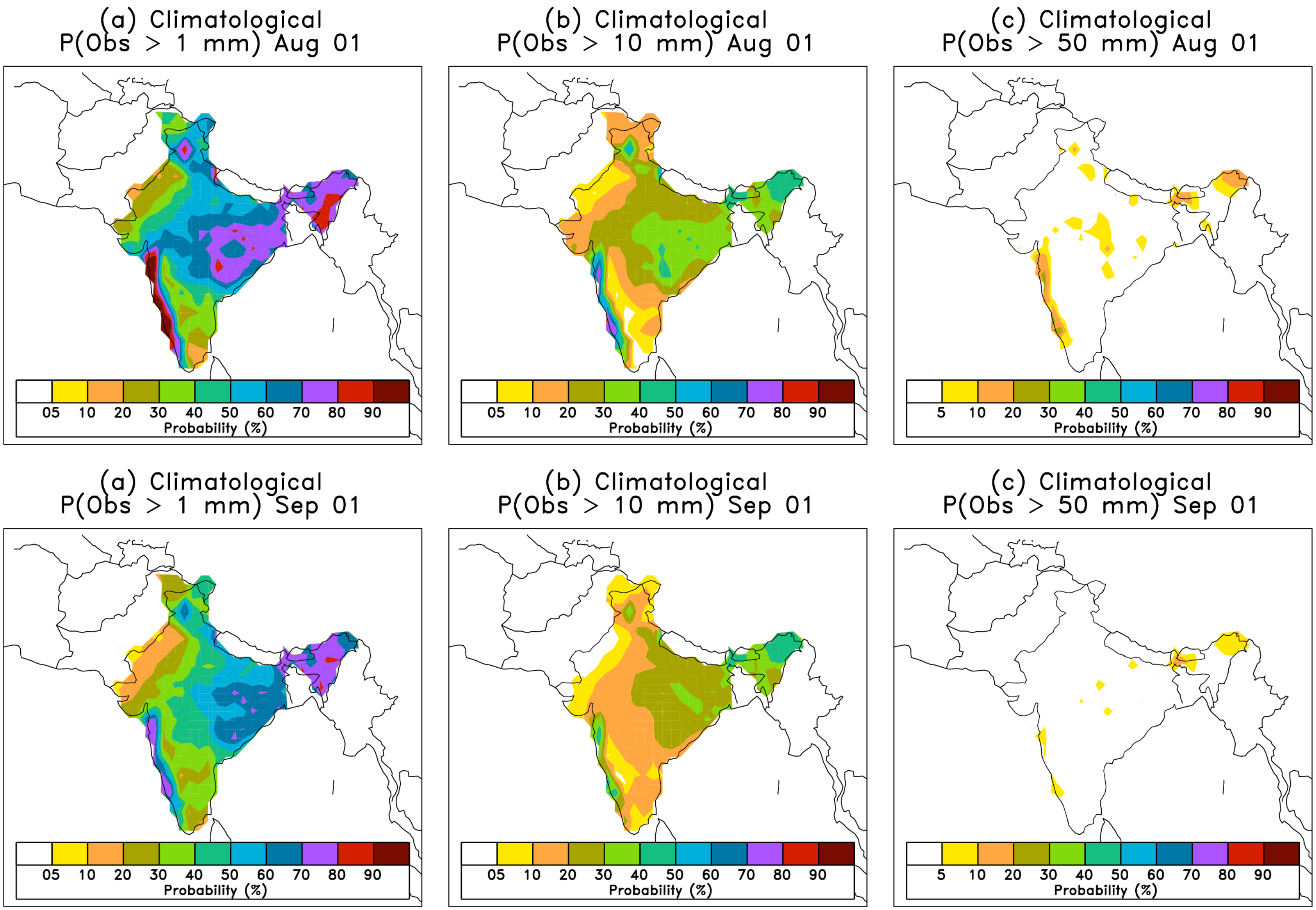
Application: monsoon forecasts over India

For this experiment we saved forecast total precipitation, column precipitable water, and sea-level pressure tendency on coarse and fine grids, as shown, for May 15 - Oct 15, 1979-2007. 1-degree precipitation analyses available over India.

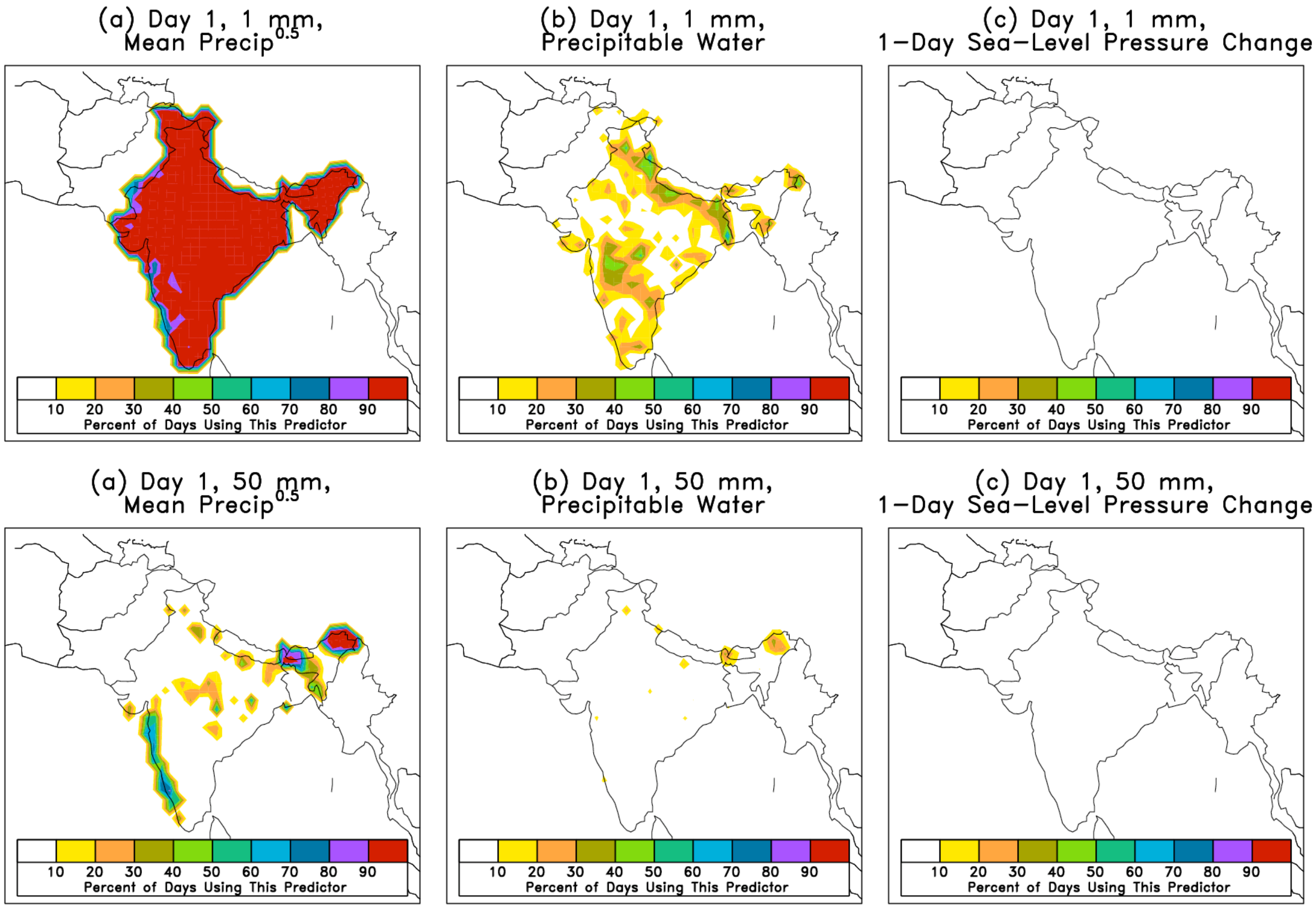
Monsoon precipitation climatology



Monsoon precipitation climatology

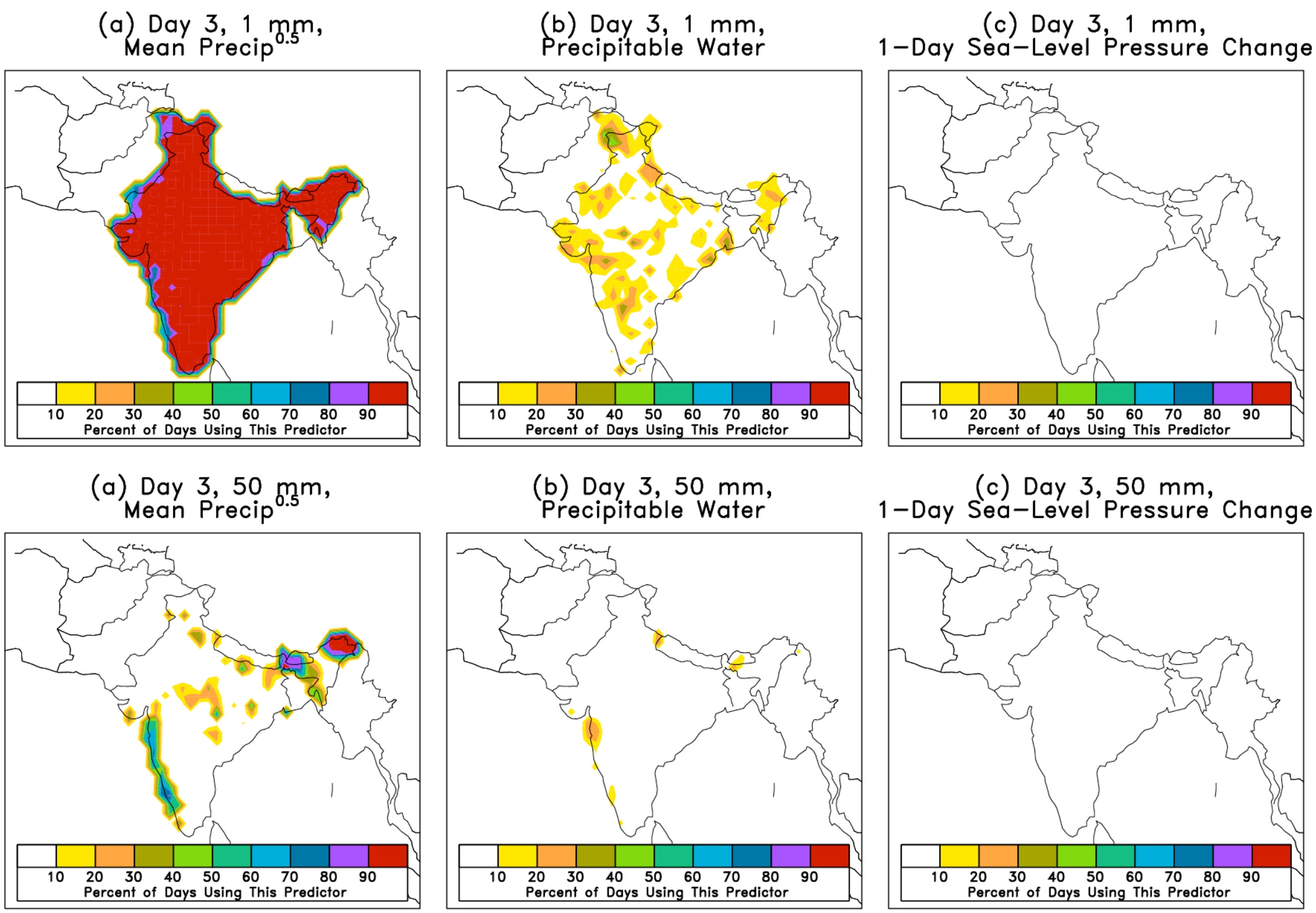


Which predictors in logistic regression with stepwise elimination? Day 1



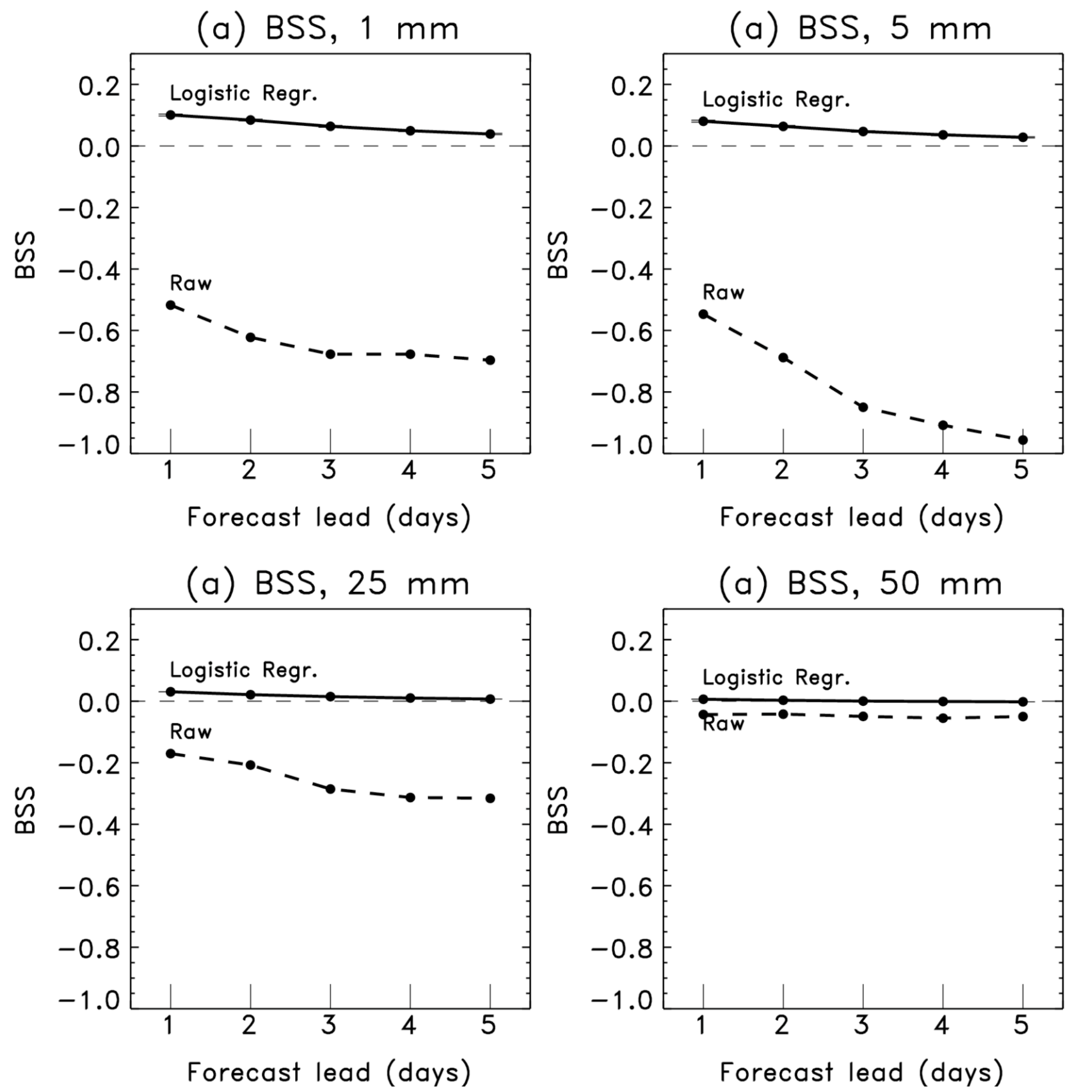
For every day of the monsoon season, a stepwise linear regression was run to determine which predictors provided a reduction in error. As shown, a power-transformed ensemble-mean forecast precipitation was uniformly selected as an important predictor. Precipitable water was occasionally selected, and sea-level pressure change was virtually never selected. Based on these results, all subsequent logistic regression analyses will be based on using only one predictor, the power-transformed ensemble-mean precipitation amount.

Which predictors in logistic regression with stepwise elimination? Day 3



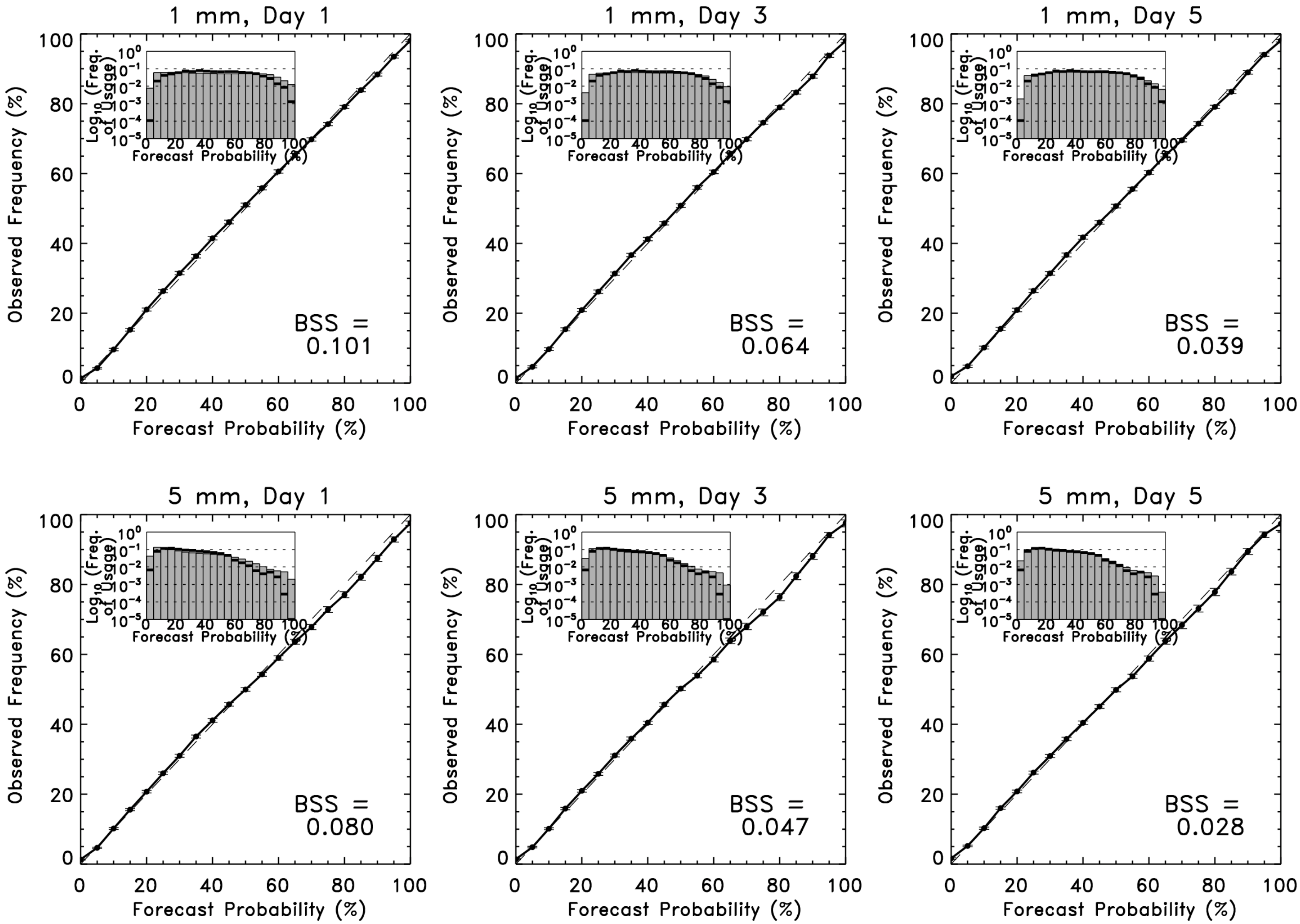
The same conclusion is reached when considering other forecast leads.

Brier Skill Scores

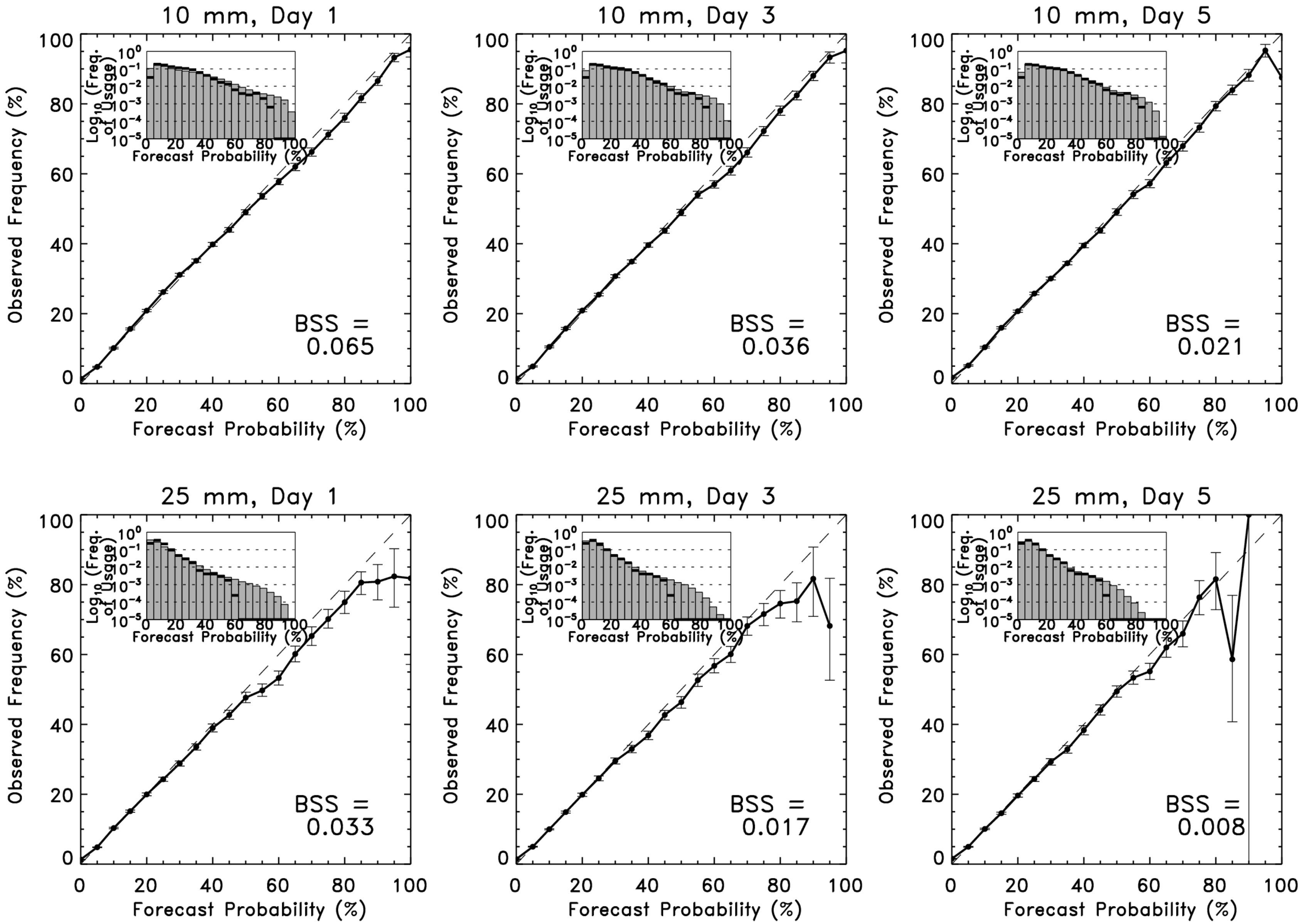


Notes:
(1) My BSS smaller than conventional BSS, since calculated in a way to not exaggerate skill.
(2) Confidence intervals are so small they don't show up on the plot.

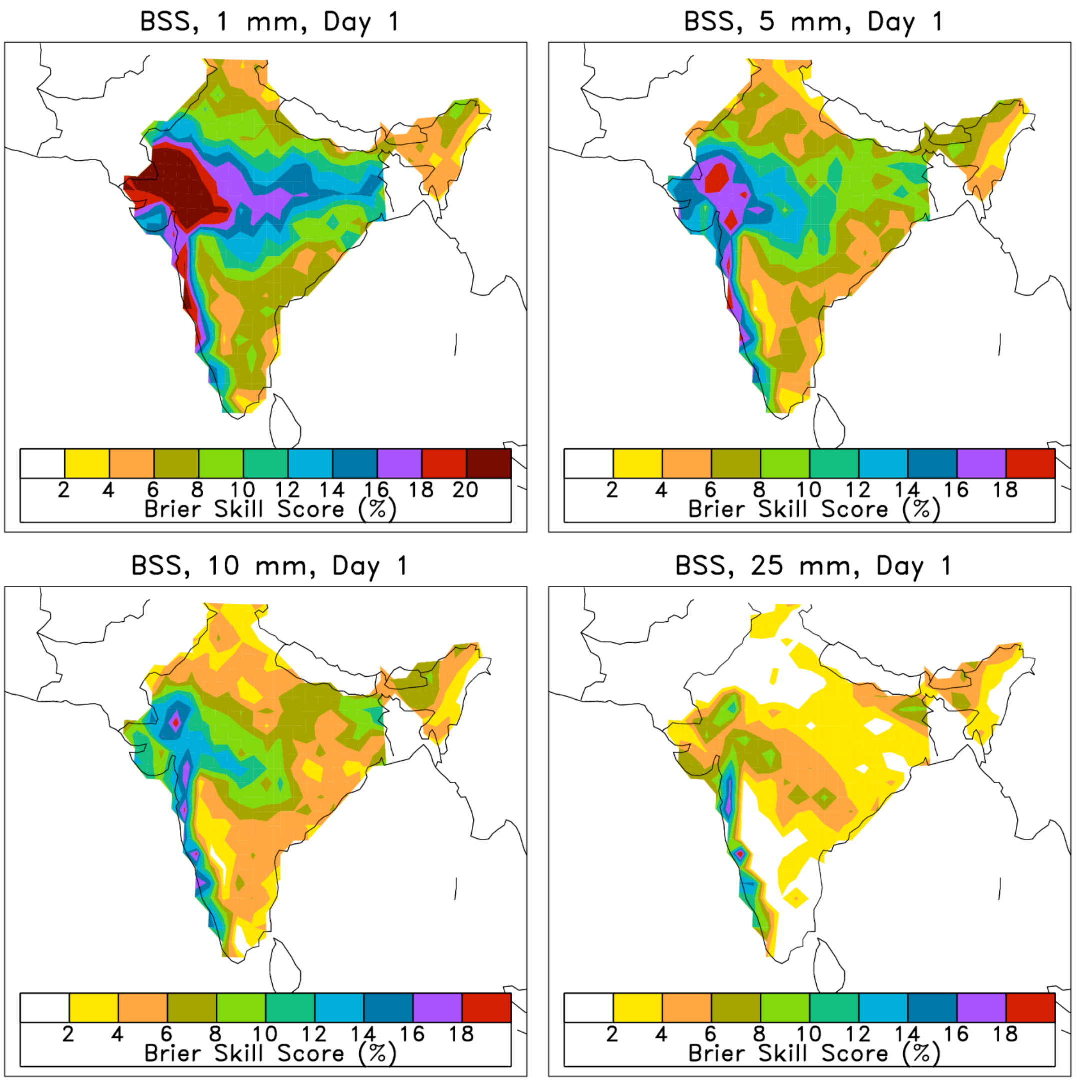
Reliability, logistic regression, 1 and 5 mm



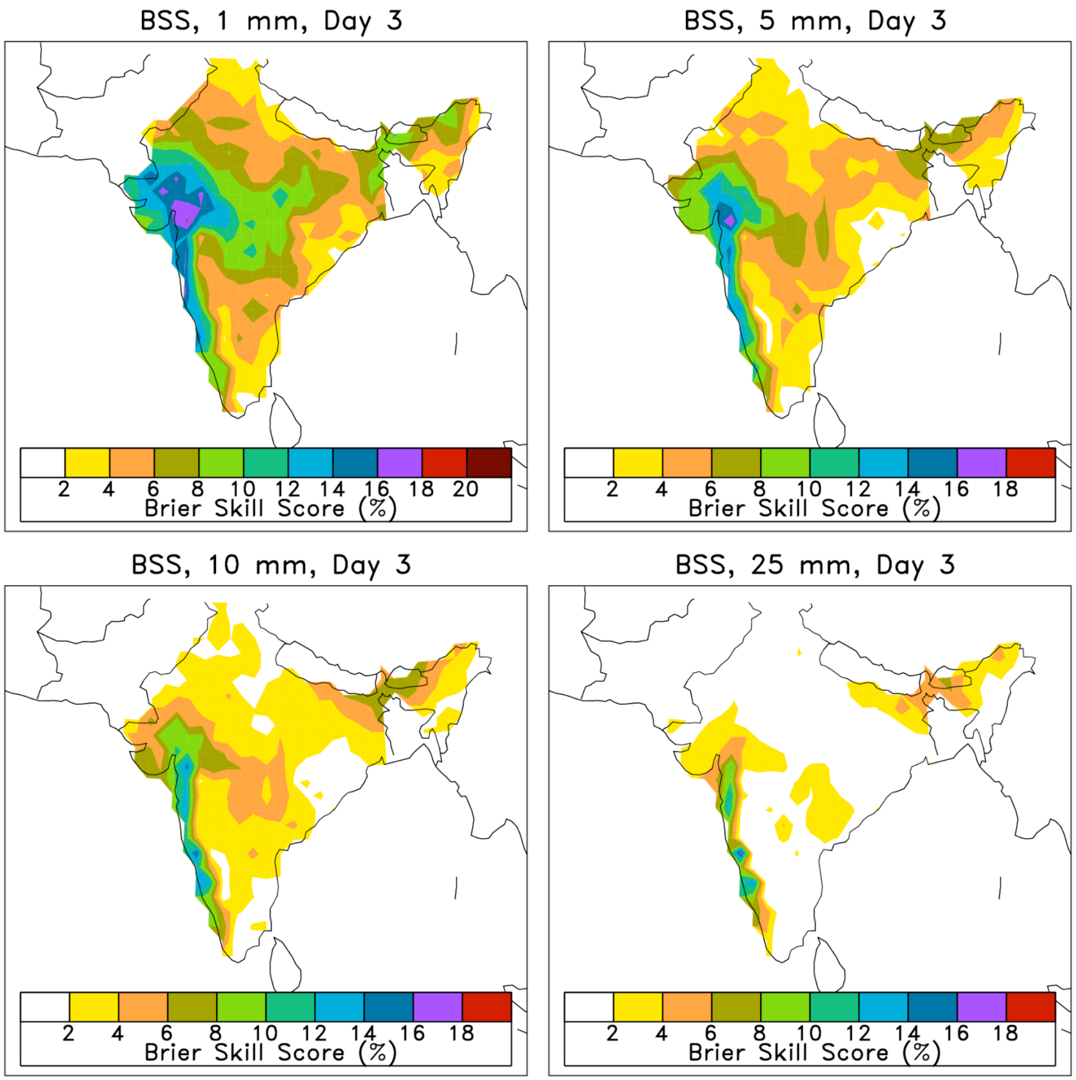
Reliability, logistic regression, 10 and 25 mm



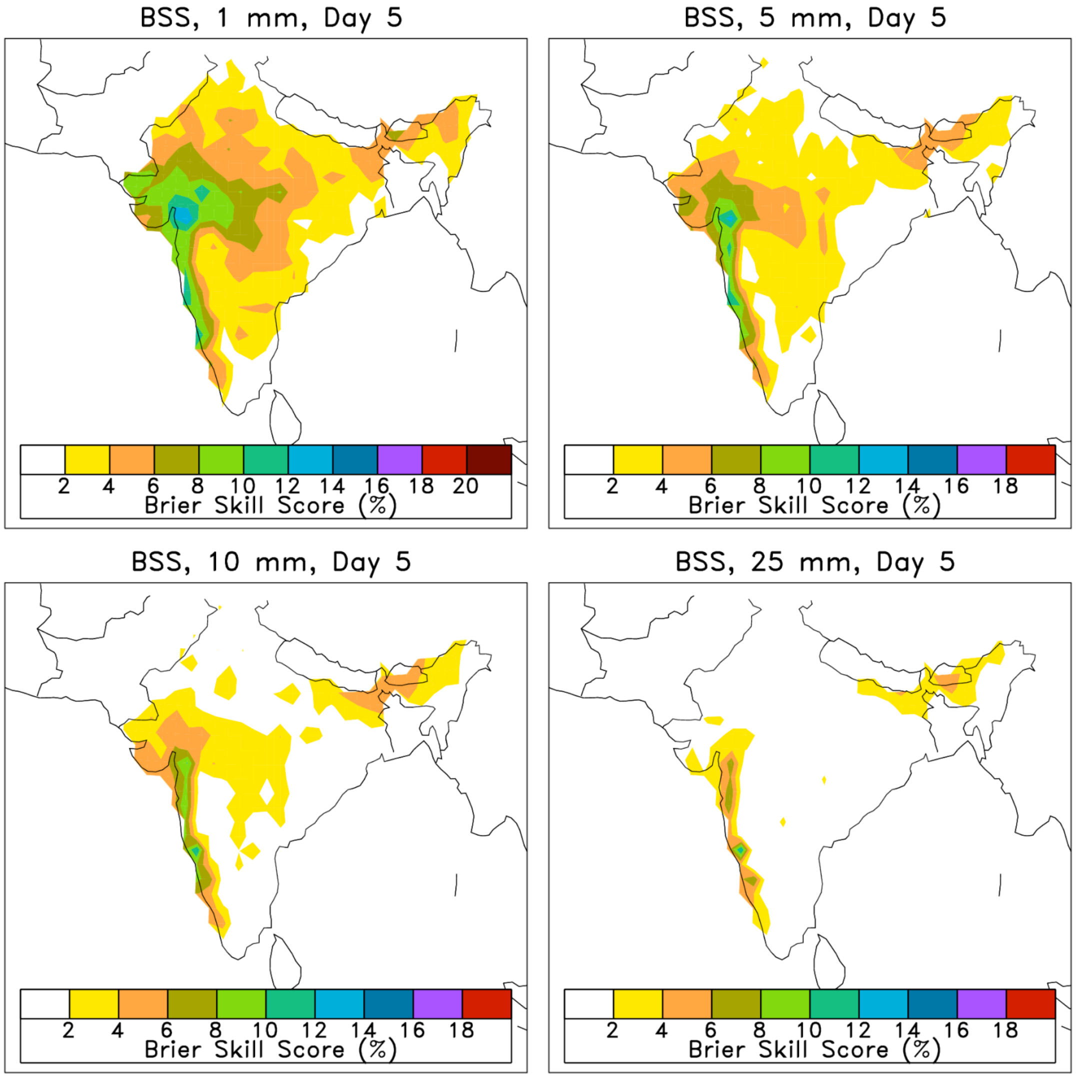
Map of logistic regression BSS, day 1



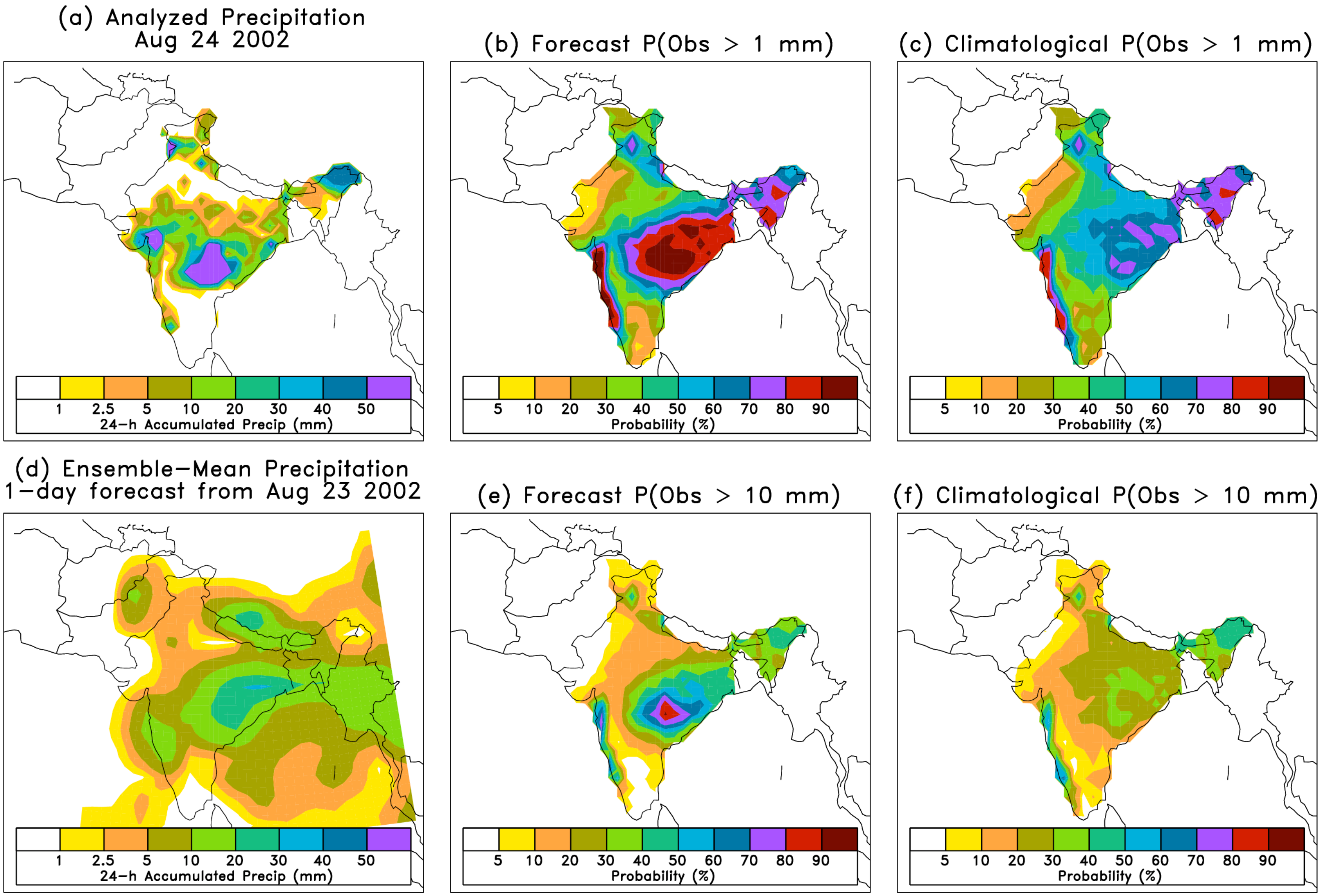
Map of logistic regression BSS, day 3



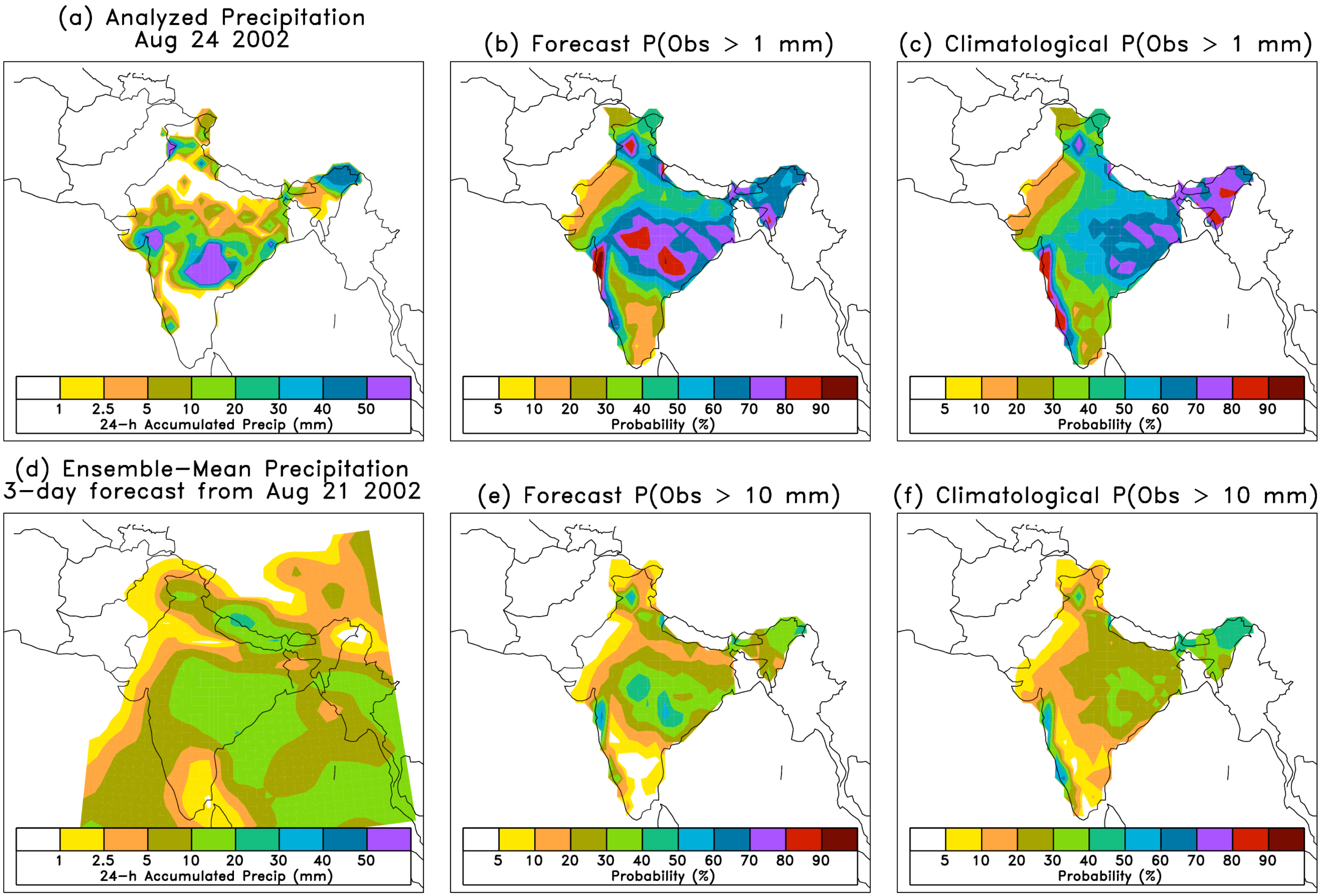
Map of logistic regression BSS, day 5



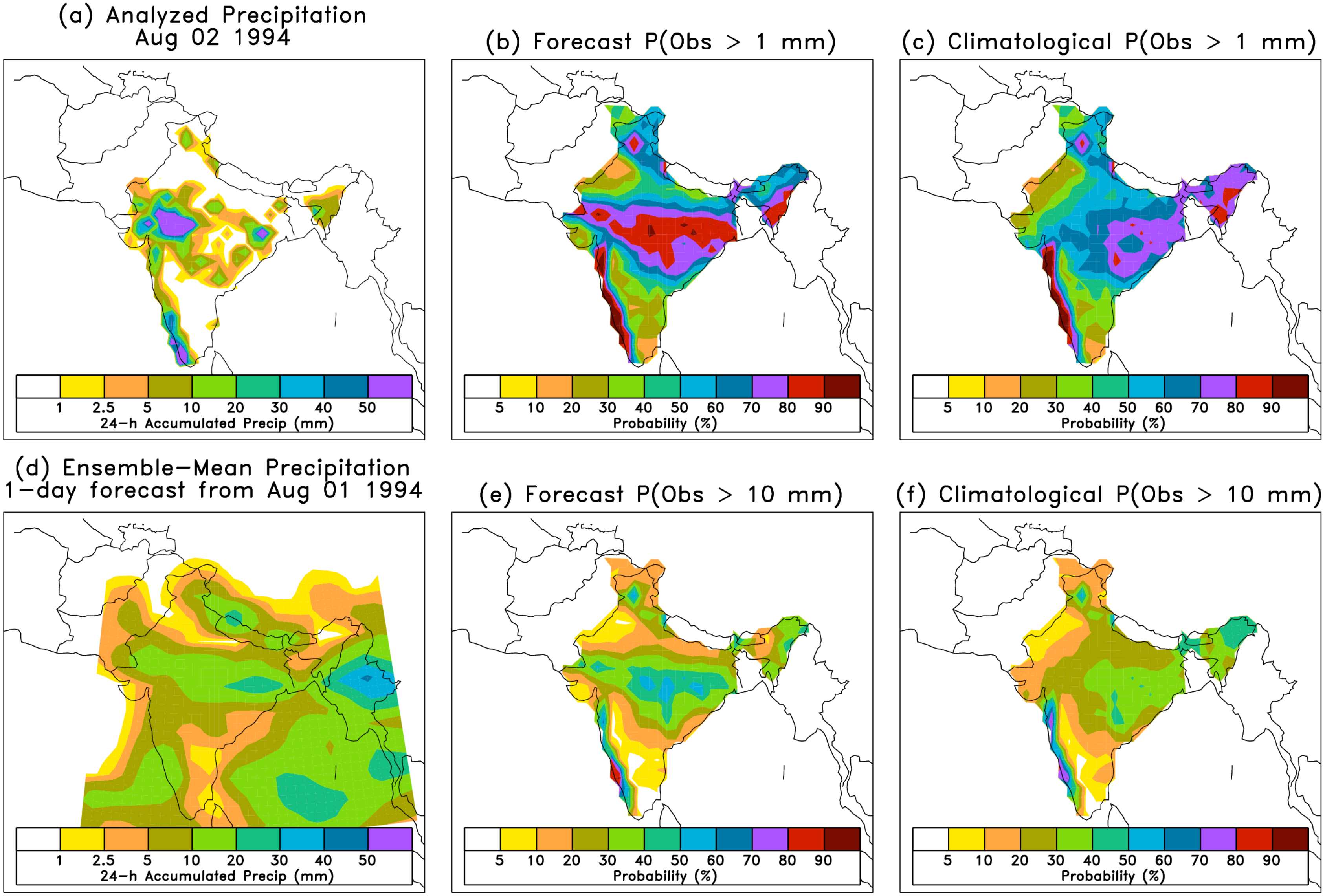
Logistic regression forecast example #1, 1-day lead



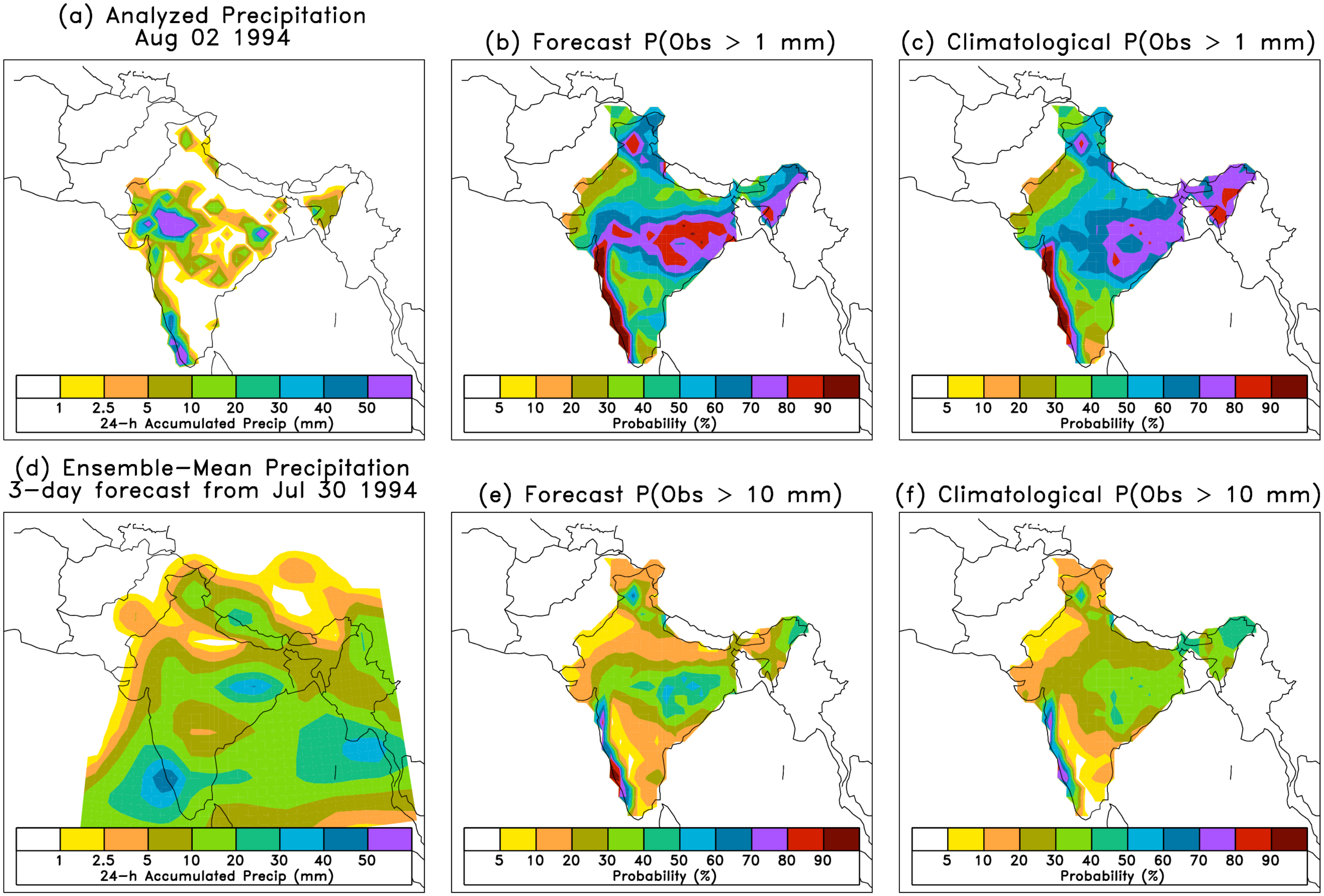
Logistic regression forecast example #1, 3-day lead



Logistic regression forecast example #2, 1-day lead

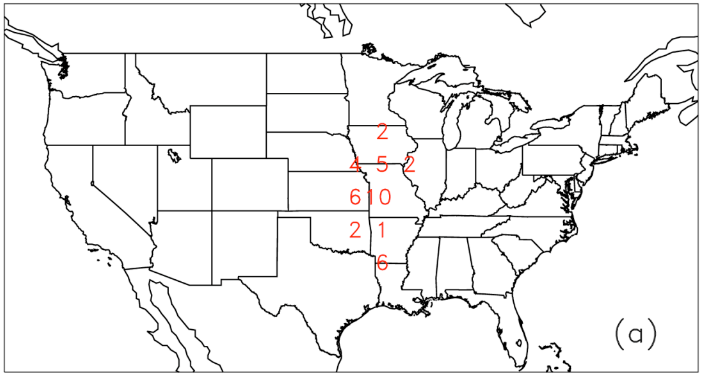


Logistic regression forecast example #2, 3-day lead

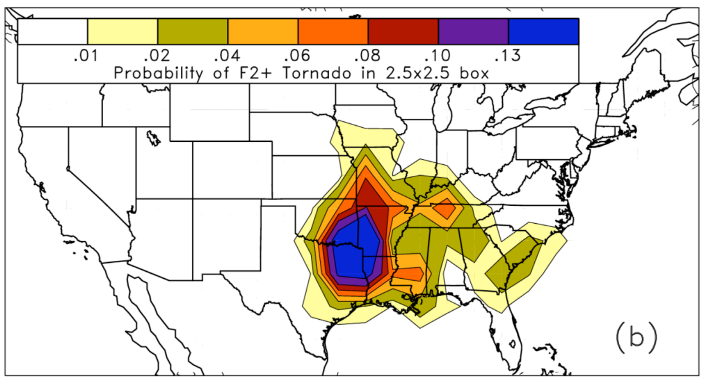


Tornado probability forecasting

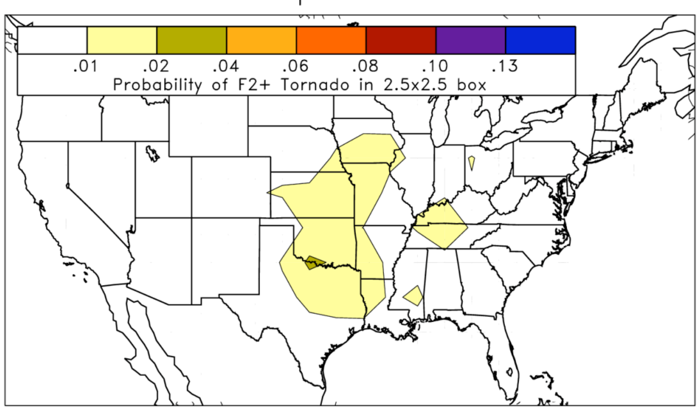
Observed F2+ Tornado Counts in 12–hour Window
Centered on 0000 UTC 27 Apr 1991



Tornado Probabilities for
01–day Forecast from 26 Apr 1991

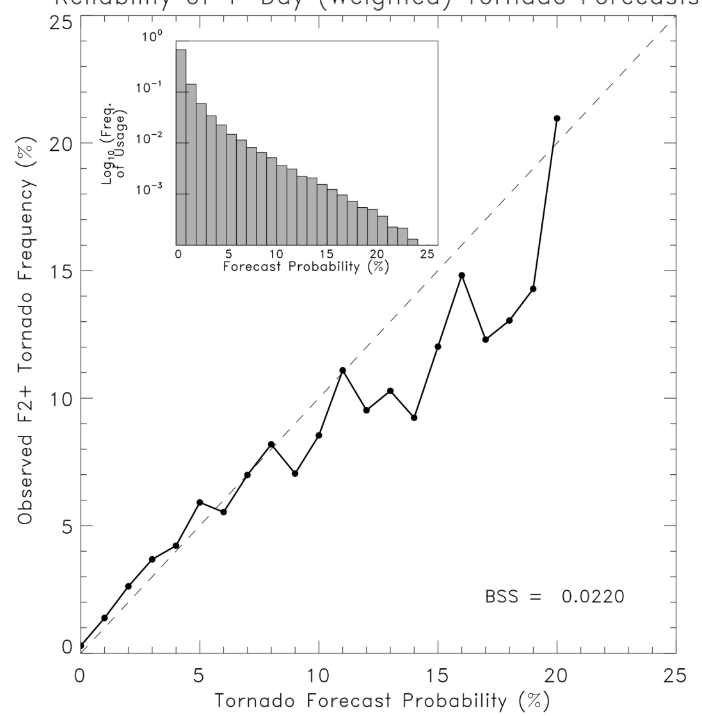


Climatological F2+ Tornado Probabilities,
15 Apr – 15 Jun



forecast wind shear and instability were used as predictors in an analog approach.

Reliability of 1–Day (Weighted) Tornado Forecasts



Part II:

Calibration using ECMWF reforecast data set

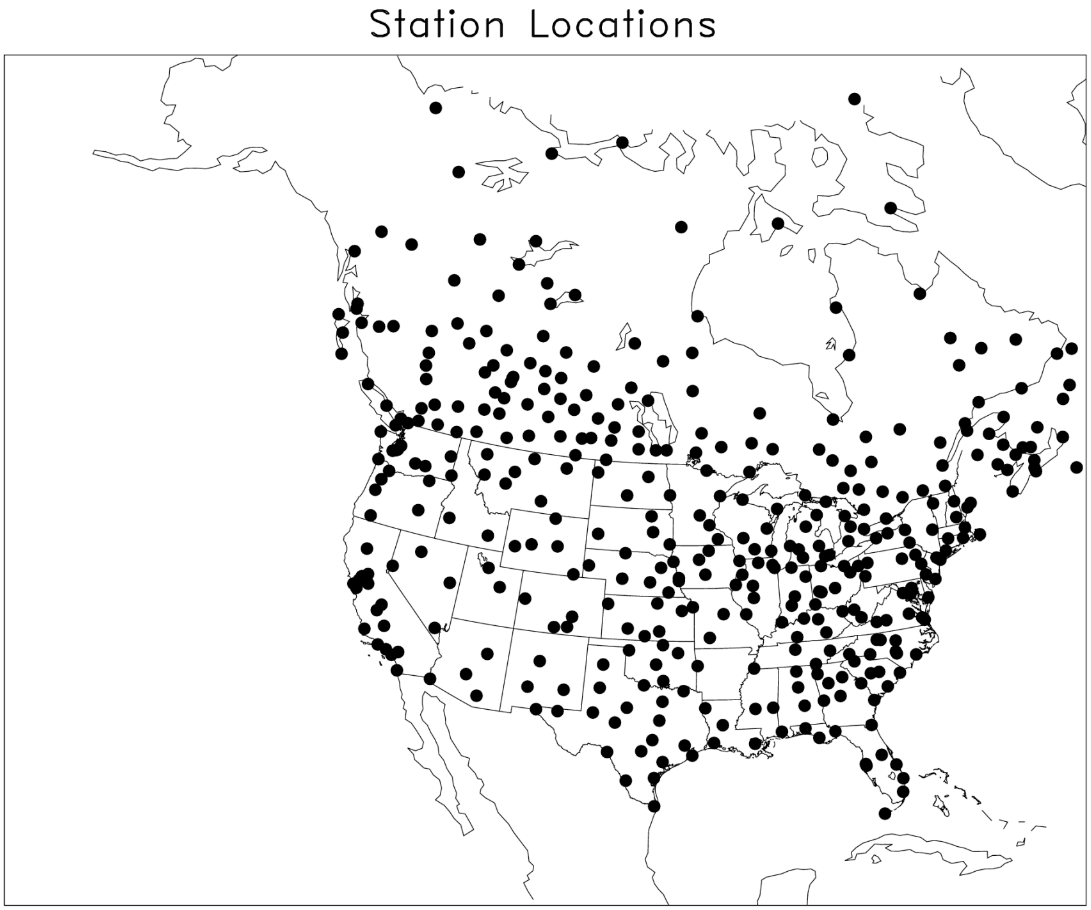
Questions

- Will reforecasts benefit calibration of a state-of-the art model like ECMWF's as much as with now outdated GFS model?
- How do probabilistic forecasts from the old GFS, with calibration, compare to the new ECMWF without?
- Are multi-decadal reforecasts really necessary? Given the computational expense of computing them, are much smaller training data sets adequate for probabilistic forecast calibration?

ECMWF's reforecast data set

- **Model**: 2005 version of ECMWF model; T255 resolution.
- **Initial Conditions**: 15 members, ERA-40 analysis + singular vectors
- **Dates of reforecasts**: 1982-2001, Once-weekly reforecasts from 01 Sep - 01 Dec, 14 weeks total. So, 20y $\times$ 14w ensemble reforecasts = 280 samples.
- **Data** obtained by NOAA / ESRL : T_{2M} and precipitation ensemble over most of North America, excluding Alaska. Saved on 1-degree lat / lon grid. Forecasts to 10 days lead.

Observation locations for temperature calibration



Produce probabilistic forecasts at stations.

Use stations from NCAR's DS472.0 database that have more than 96% of the yearly records available, and overlap with the domain that ECMWF sent us.

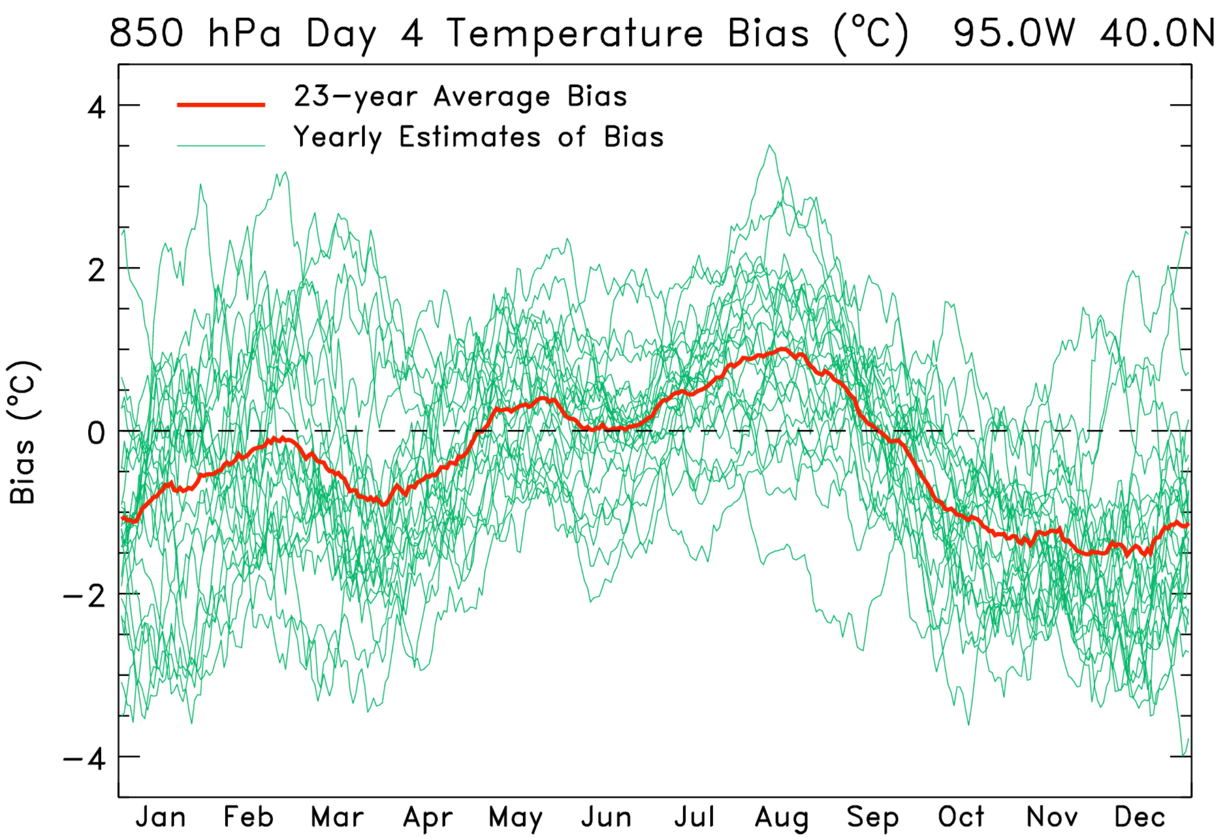
Calibration procedure: “NGR”
“Non-homogeneous Gaussian Regression”

- **Reference**: Gneiting et al., *MWR*, **133**, p. 1098. Shown in Wilks and Hamill (MWR, 135, p 2379) to be best of common calibration methods for surface temperature using reforecasts.
- **Predictors**: ensemble mean and ensemble spread
- **Output**: mean, spread of calibrated normal distribution

$$f^{CAL}\left(\bar{\mathbf{x}},\sigma\right) \sim N\left(a + b\bar{\mathbf{x}}, c + d\sigma\right)$$

- **Advantage**: leverages possible spread/skill relationship appropriately. Large spread/skill relationship, c ≈ 0.0, *d* ≈1.0. Small, *d* ≈ 0.0
- **Disadvantage**: iterative method, slow…no reason to bother (relative to using simple linear regression) if there’s little or no spread-skill relationship.

Inter-annual variability of forecast bias

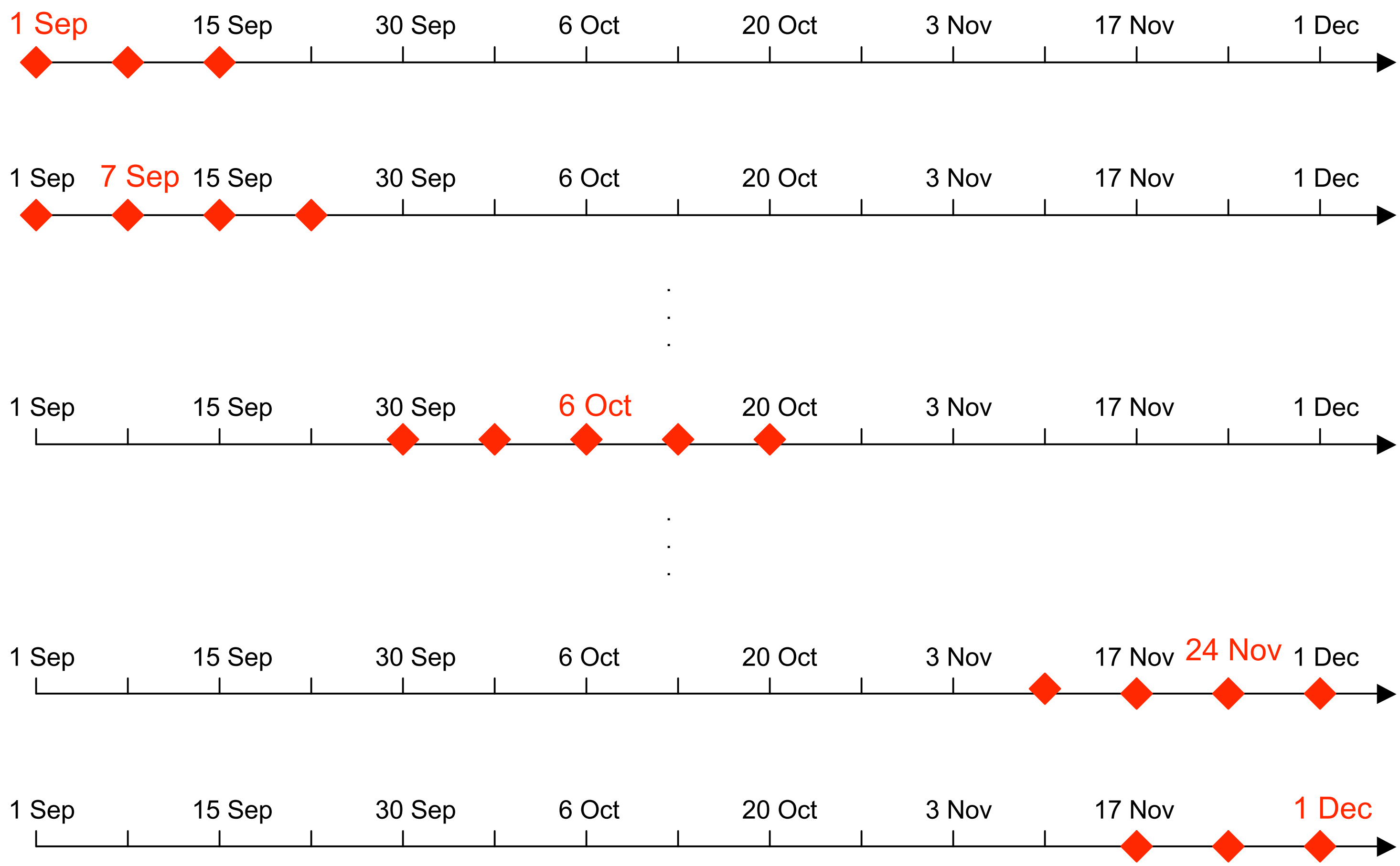


Red curve shows bias averaged over 23 years of data (bias = mean F-O in running 61-day window)

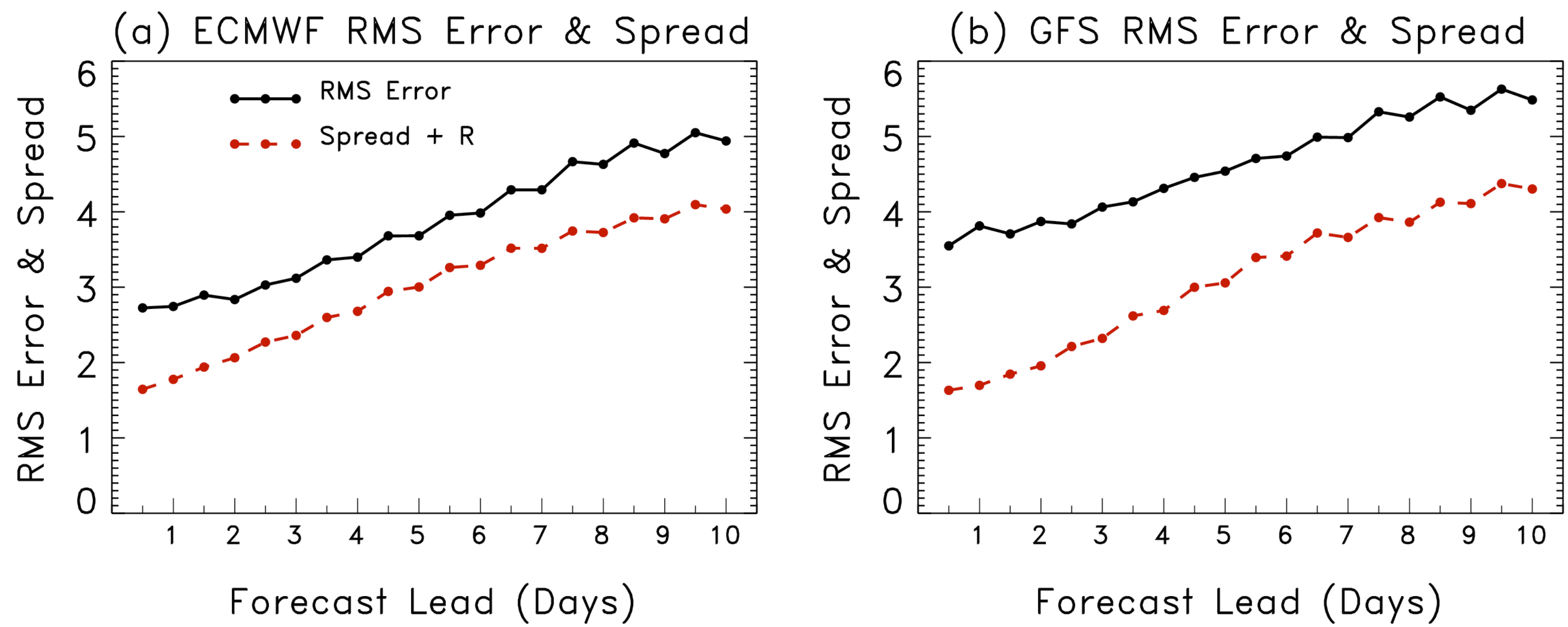
Green curves show 23 individual yearly running-mean bias estimates

Note large inter-annual variability of bias.

What training data to use, given inter-annual variability of bias?

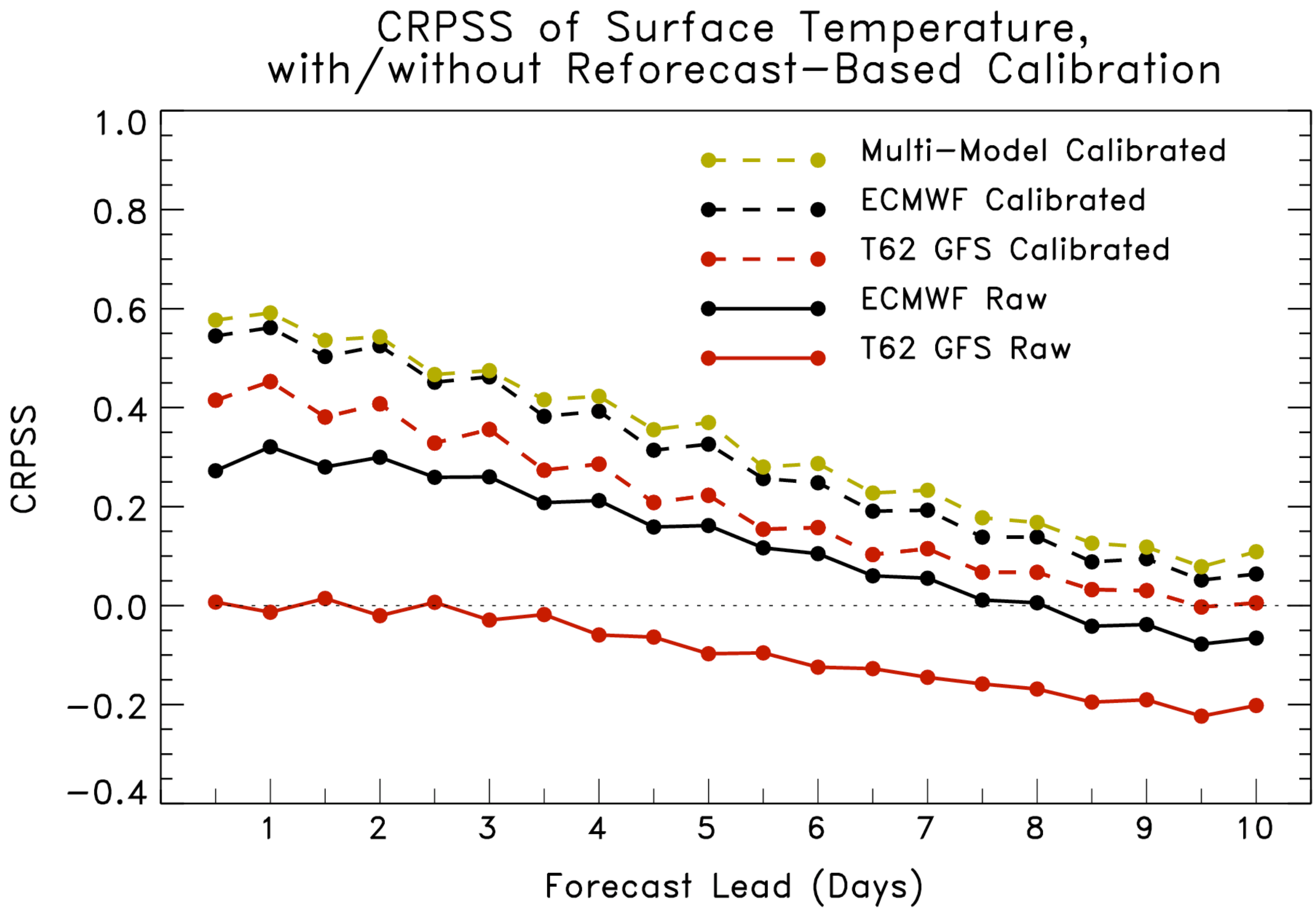


Forecast spread and error



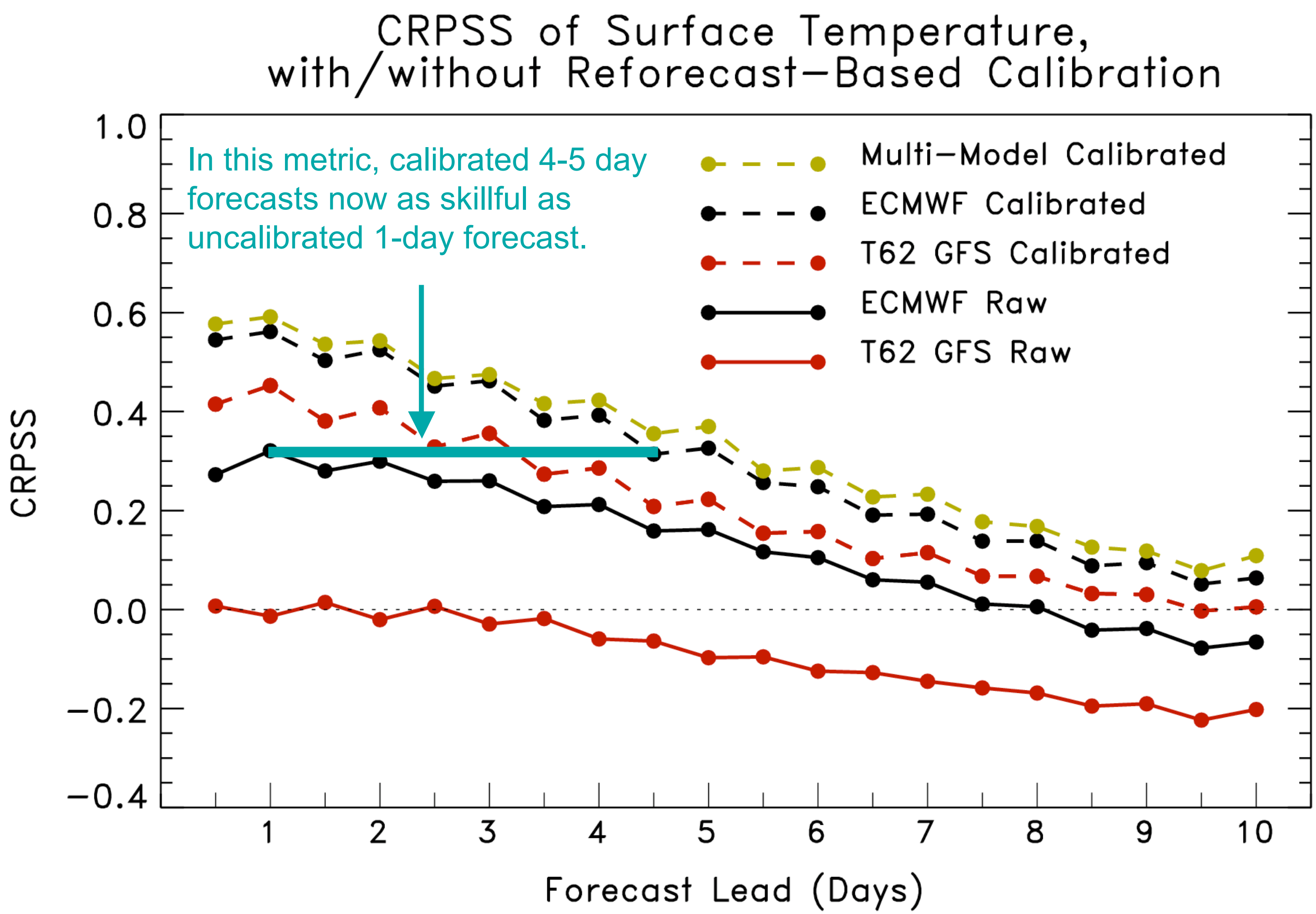
For both systems, with 2-m temperature, there is a deficiency of spread. This is much worse for GFS than ECMWF.

ECMWF, raw and post-processed



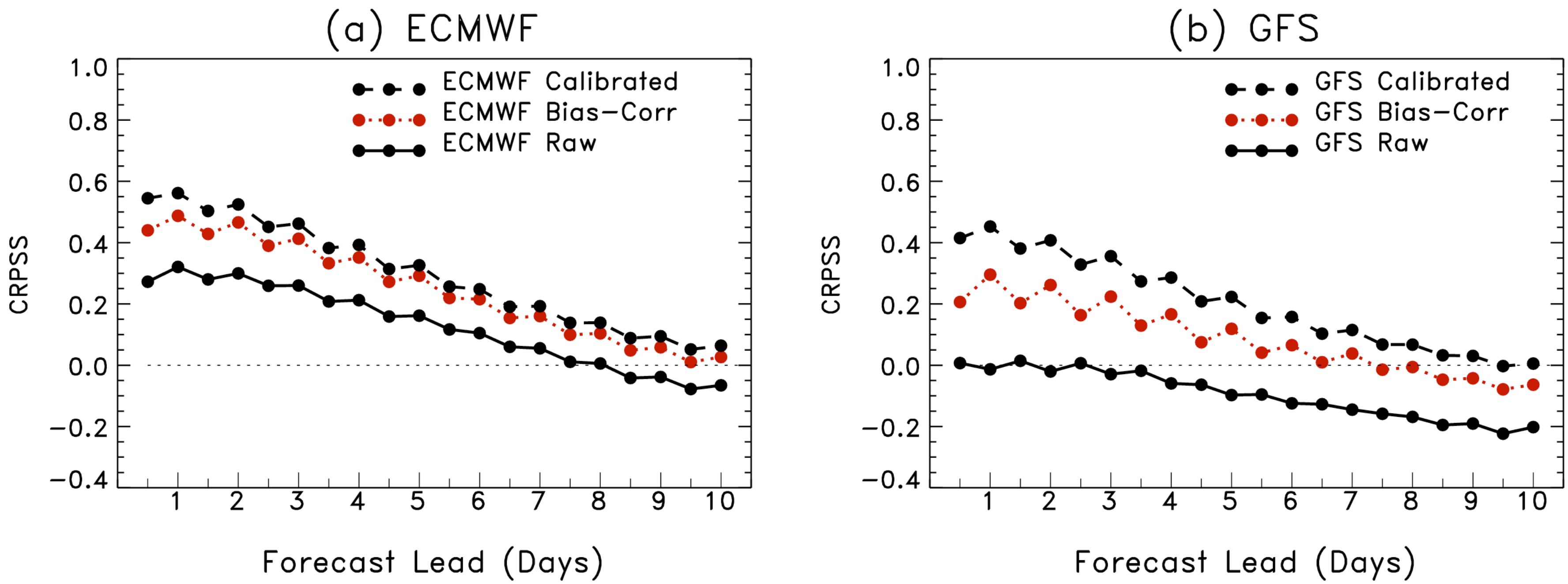
Note: 5th and 95th %ile confidence intervals very small, 0.02 or less

ECMWF, raw and post-processed



Note: 5th and 95th %ile confidence intervals very small, 0.02 or less

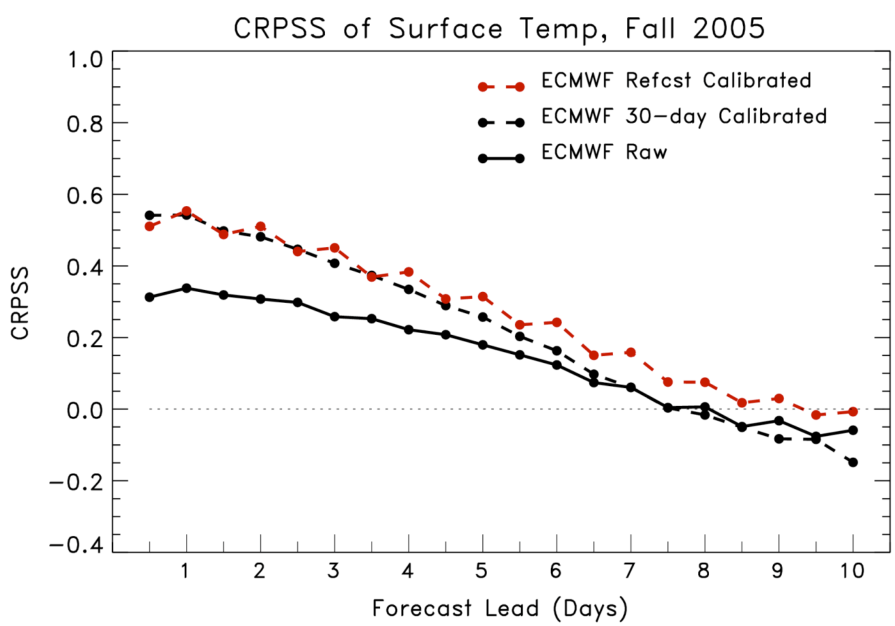
How much from simple bias correction?



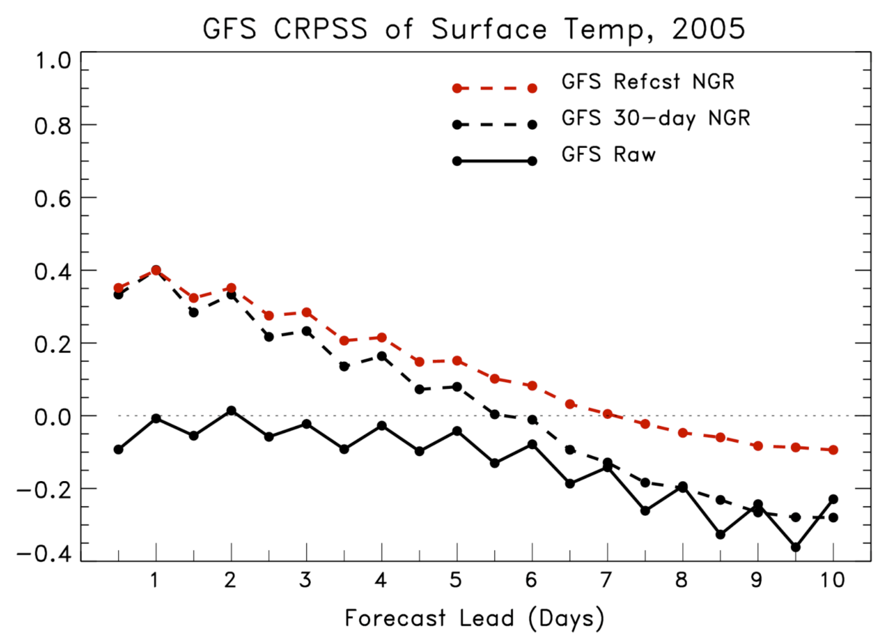
~ 60 percent of total improvement at short leads, 70 percent at longer leads.

How much from short training data sets?

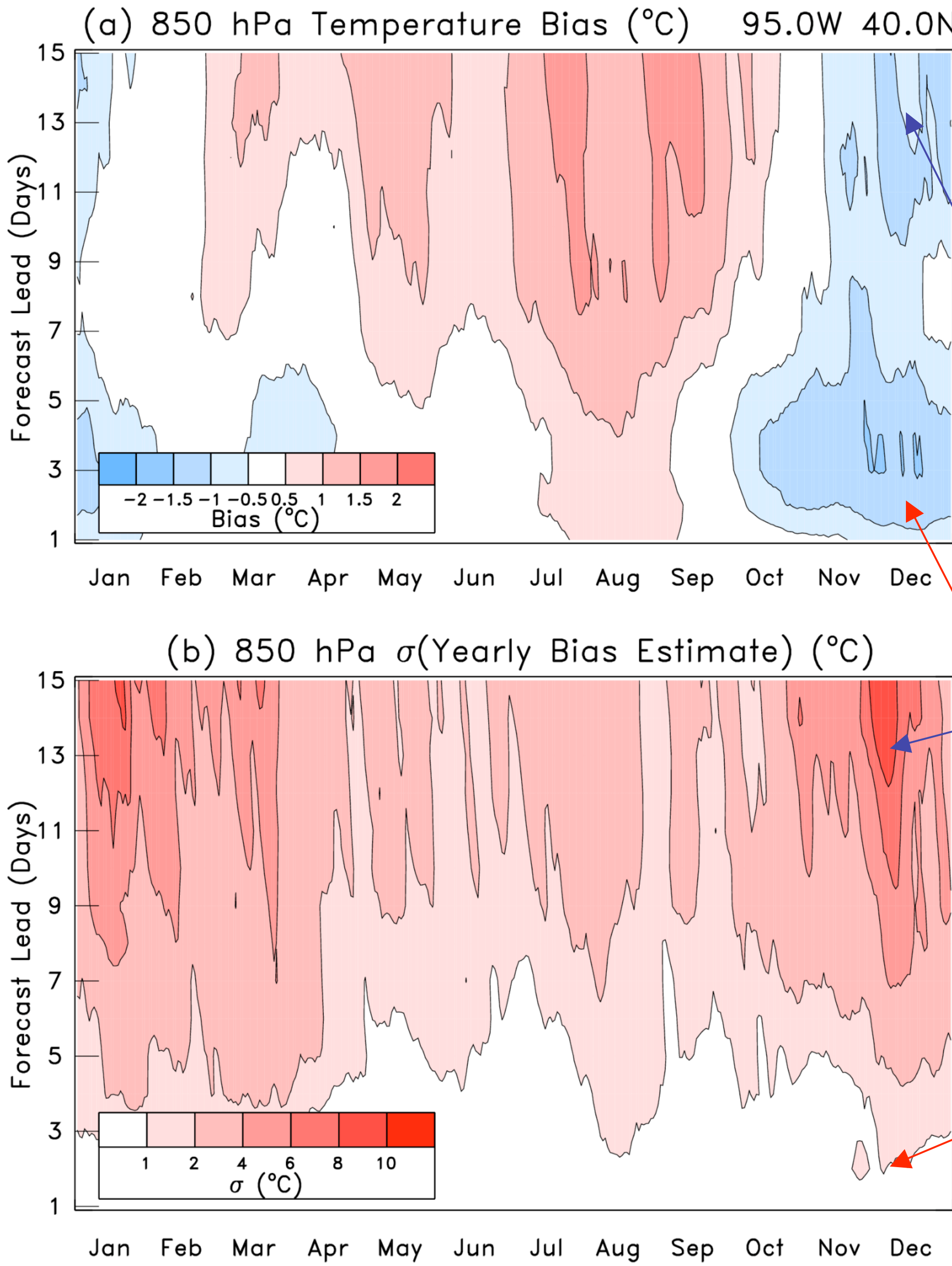
ECMWF



GFS



Note: (1) that ECMWF reforecasts use 3D-Var initial condition, 2005 real-time forecasts use 4D-Var. This difference may lower skill with reforecast training data set. (2) No other predictors besides forecast T2m; perhaps with, say, soil moisture as additional predictor, reforecast calibration would improve relative to 30-day.



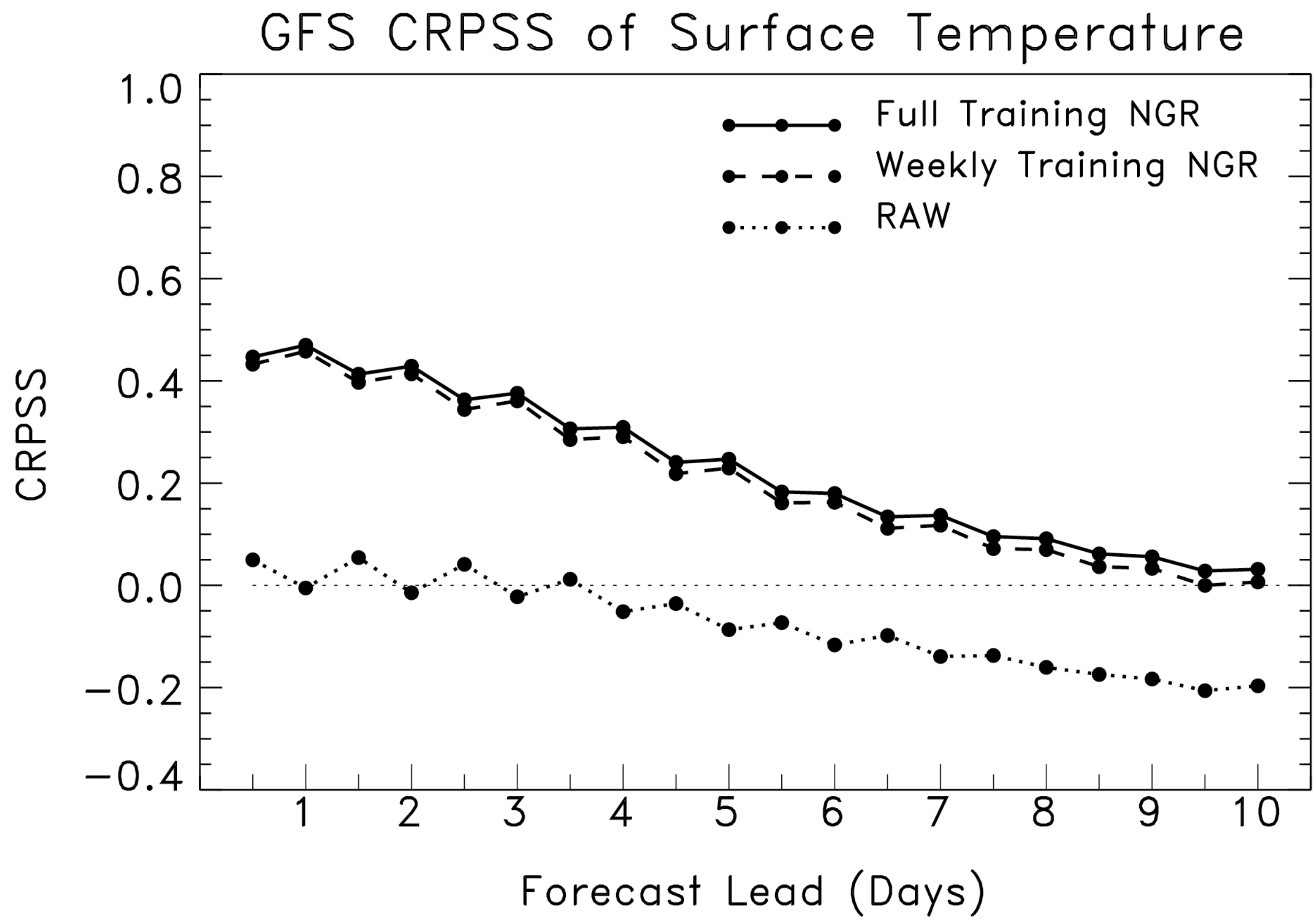
When are long reforecast data sets necessary, and when are they not?

Example: bias correction.

Here, **large training data set required**; bias is small relative to its yearly variability.

Here, **small training data set adequate**; bias comparable or greater than its yearly variability.

How much from long GFS training data set?



Here GFS reforecasts sampled once per week are compared to those sampled once per day ("full").

Precipitation calibration

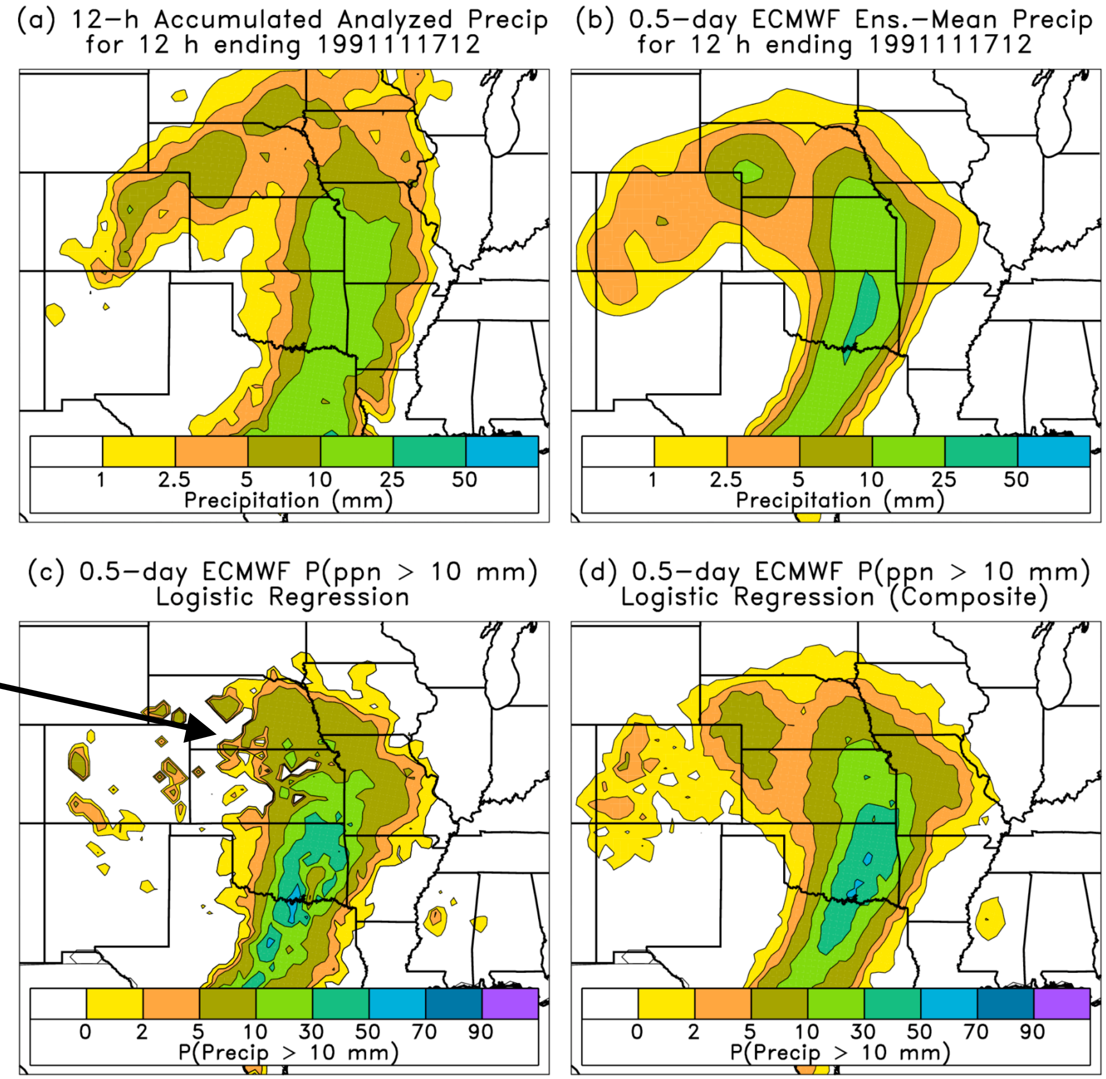
- NARR CONUS 12-hourly data used for training, verification. ~32 km grid spacing
- Logistic regression for calibration here

$$P(O > T) = 1.0 - \frac{1.0}{1.0 + \exp\left\{\beta_0 + \beta_1 \left(\overline{x}^f\right)^{0.25} + \beta_2 \left(\sigma^f\right)^{0.25}\right\}}$$

- More weight to samples with heavier forecast precipitation to improve calibration for heavy-rain events.
- Unlike temperature, throw Sep-Dec training data together.

Problem: patchy probabilities when grid point X trained with only grid point X's forecasts / obs

Even 20 years of weekly forecast data (260 samples after cross-validation) is not enough for stable regression coefficients, especially at higher precipitation thresholds.



When is it proper to use training data at location B to supplement regression analysis at location A?

(1) When location B's errors are independent of location A's errors.

(2) When observed CDF at A and B are very similar.

(3) When forecast CDF at A and B are very similar.

(4) When corr(forecast, observed) at A and B are similar.

When is it proper to use training data at location B to supplement regression analysis at location A?

(1) When location B's errors are independent of location A's errors.

Make sure location A is not too close to location B

(2) When observed CDF at A and B are very similar.

(3) When forecast CDF at A and B are very similar.

(4) When corr(forecast, observed) at A and B are similar.

When is it proper to use training data at location B to supplement regression analysis at location A?

(1) When location B’s errors are independent of location A’s errors.

(2) When observed CDF at A and B are very similar.

Need lots of samples. Luckily, ~28 year NARR provides them.

(3) When forecast CDF at A and B are very similar.

(4) When corr(forecast, observed) at A and B are similar.

When is it proper to use training data at location B to supplement regression analysis at location A?

(1) When location B's errors are independent of location A's errors.

(2) When observed CDF at A and B are very similar

(3) When forecast CDF at A and B are very similar.

(4) When corr(forecast, observed) at A and B are similar.

Judging this would be tough with ECMWF forecasts. Only 14 weeks*20 years, not a large sample for non-normally distributed data. Can be fooled by rare events.

When is it proper to use training data at location B to supplement regression analysis at location A?

(1) When location B's errors are independent of location A's errors.

(2) When observed CDF at A and B are very similar

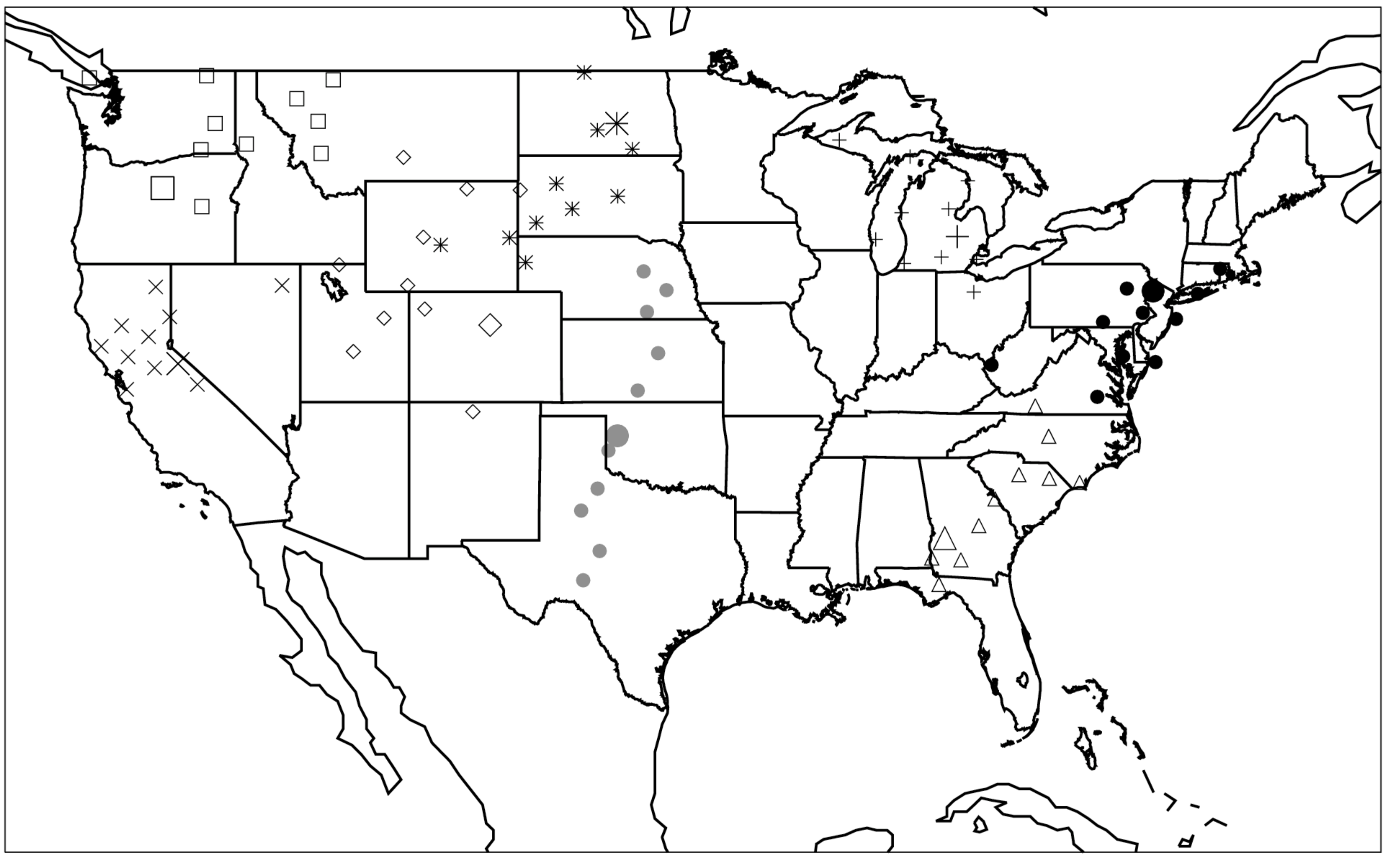
(3) When forecast CDF at A and B are very similar.

(4) When corr(forecast, observed) at A and B are similar.

Tricky to compute in dry regions, where overwhelming bulk of the samples are zero's.

Tested method: add in training data at other grid points that have similar analyzed climatologies

Selected Analog Composite Locations



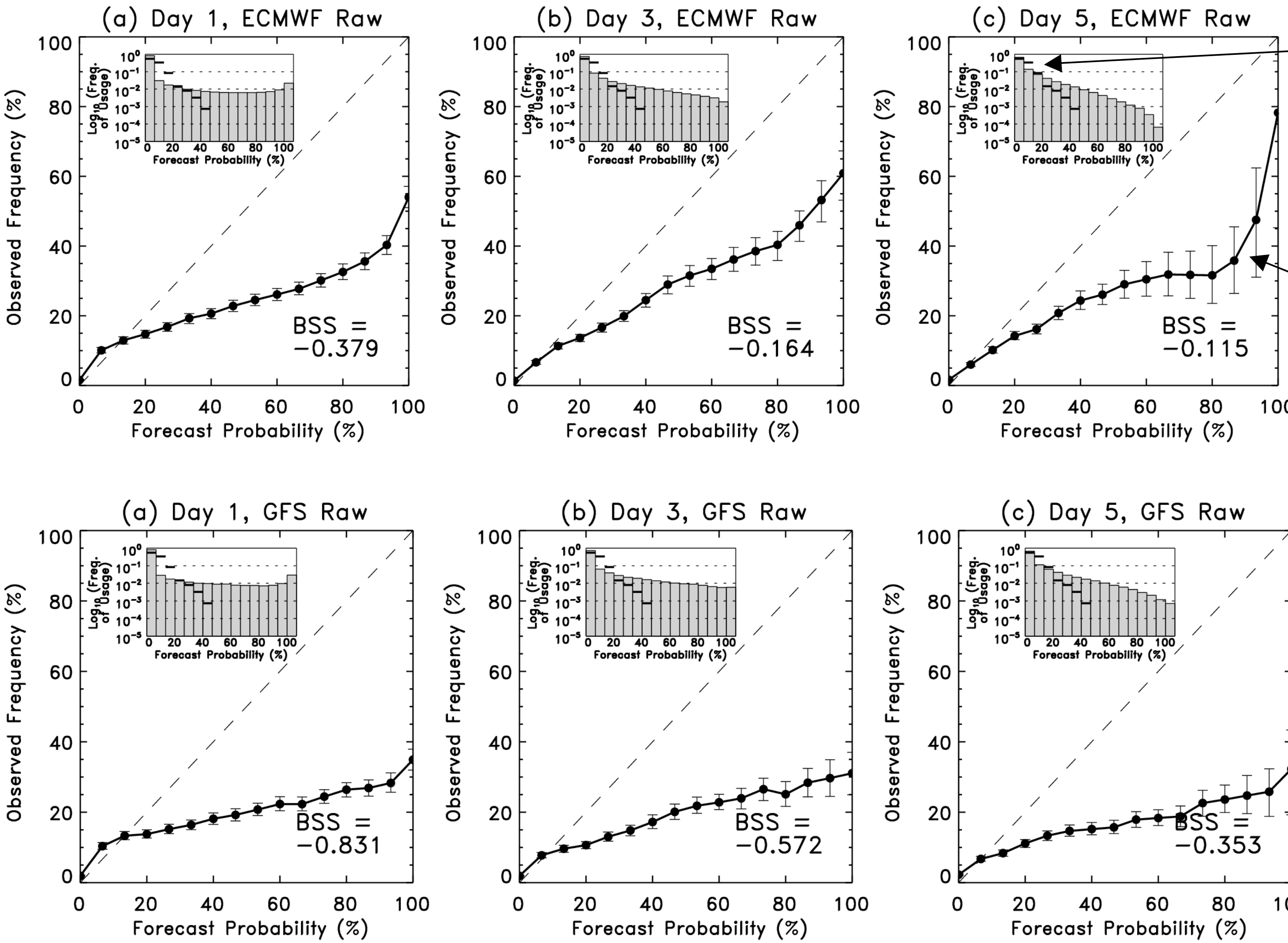
Big symbol:
grid point
where we
do regression

Small symbols:
analog locations
with similar
climatologies

Training data sets tested

- "Weekly" - use 1x weekly, 20-year reforecasts for training data. Sep-Dec cases all thrown together. X-validated.
- "30-day" - for 2005 only, where forecasts available every day, train using the prior available 30 days.
- "Full" (GFS only) - use 25 years of daily reforecasts. X-validated.

5-mm reliability diagrams, raw ensembles



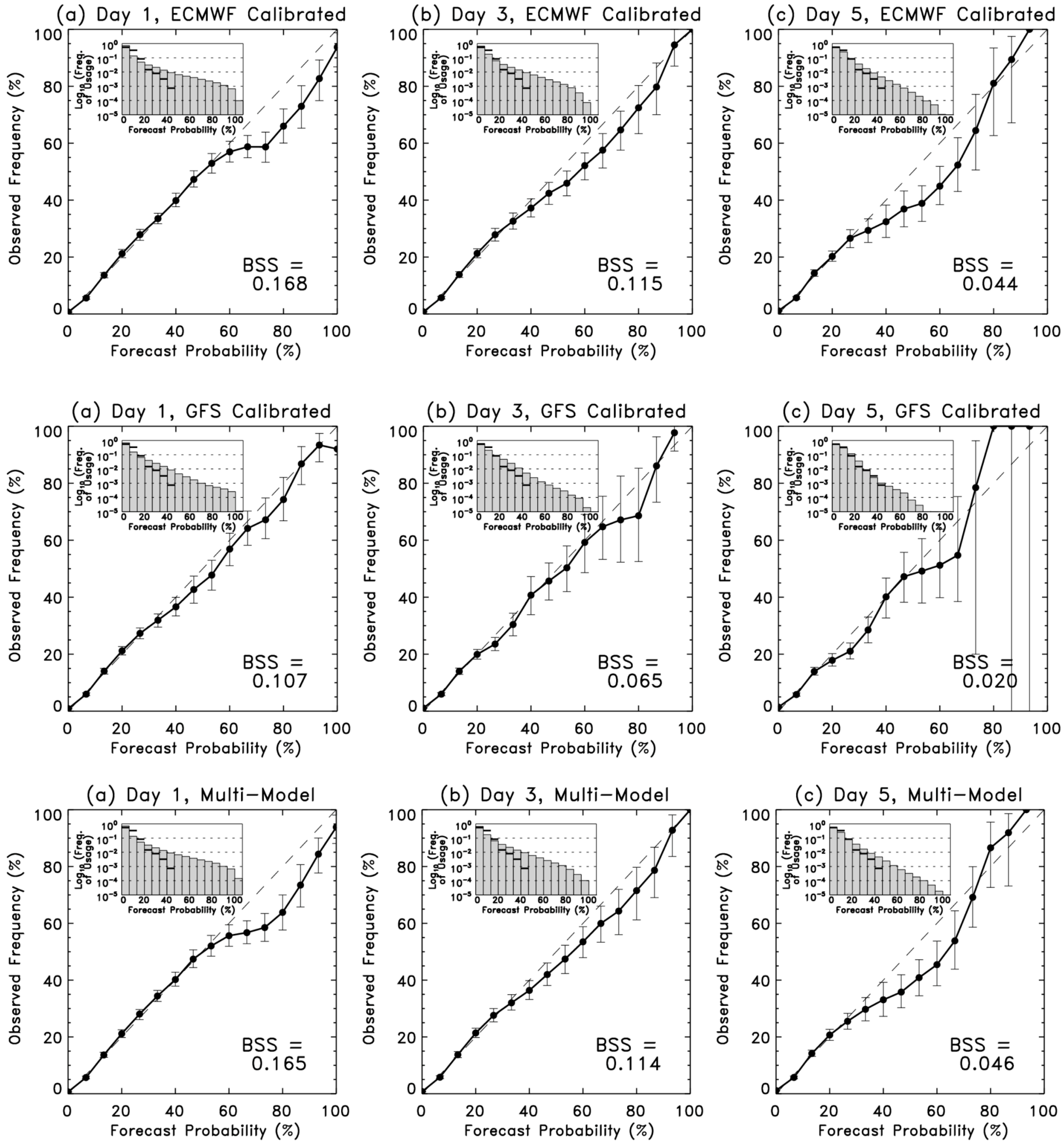
horizontal lines indicate distribution of climatology

error bars from block bootstrap

Raw forecasts have poor skill in this strict BSS

5-mm reliability diagrams, calibrated

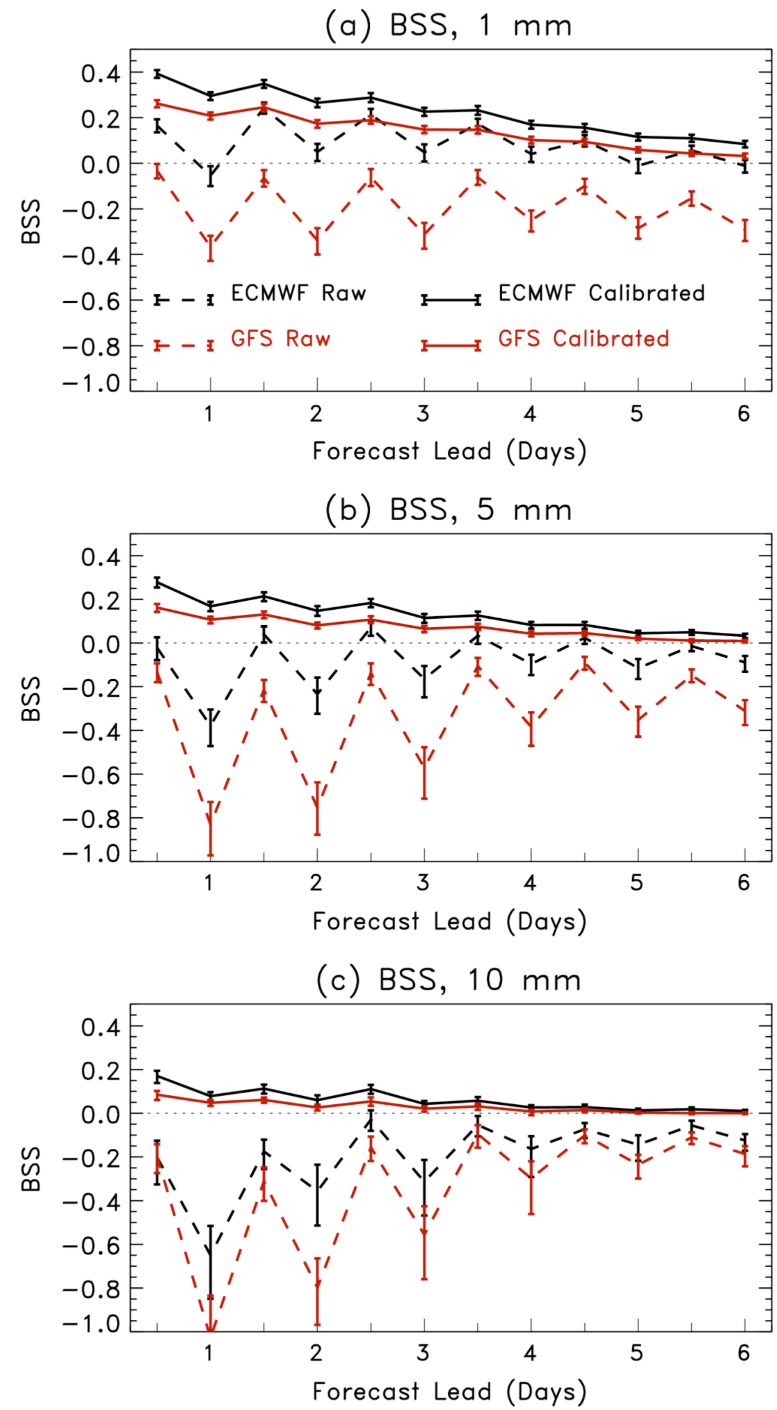
In some respects GFS forecasts look more calibrated but the frequency of usage histograms show ECMWF sharper and thus more skillful.



Brier Skill Scores

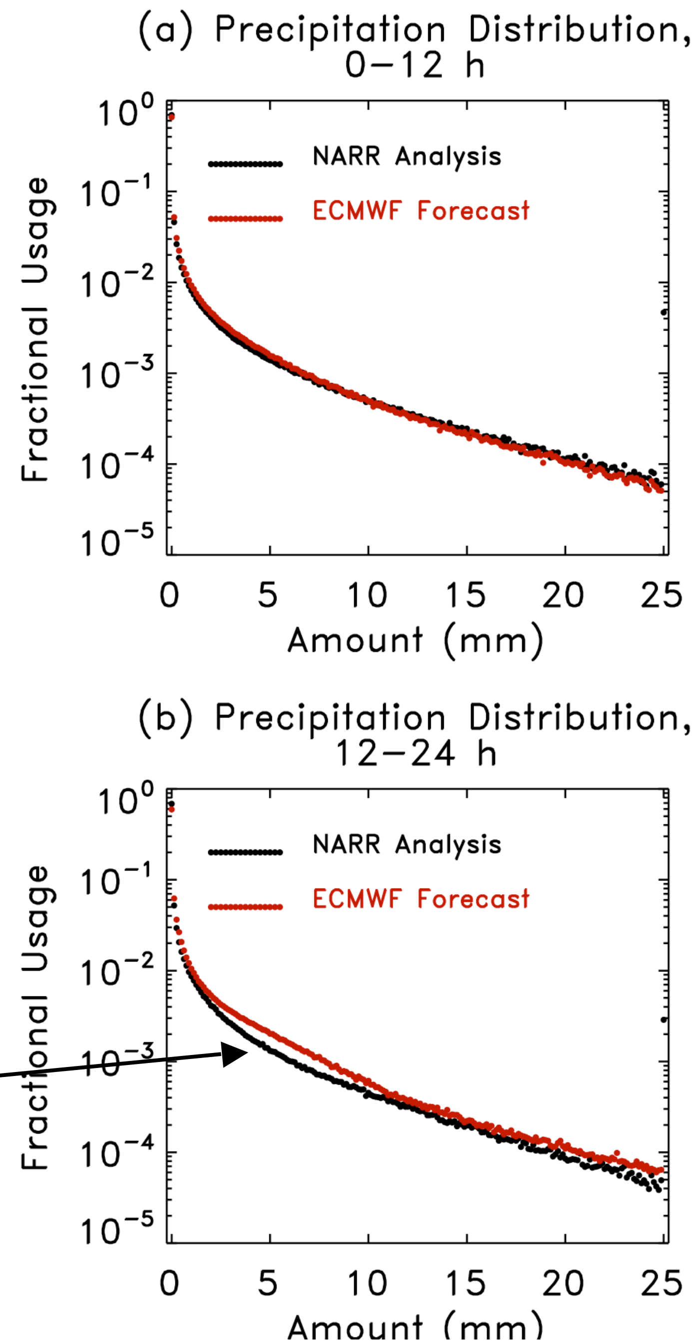
Notes:

(1) Diurnal oscillation in raw forecast skill
(2) Raw forecast skill poor, especially at higher thresholds
(3) Calibration has substantial positive impact.
(4) ECMWF > GFS skill.
(5) Multimodel not plotted, ~ same as ECMWF calibrated



Why are 12Z - 00Z forecasts less skillful?

Over-forecast bias in models during daytime relative to NARR

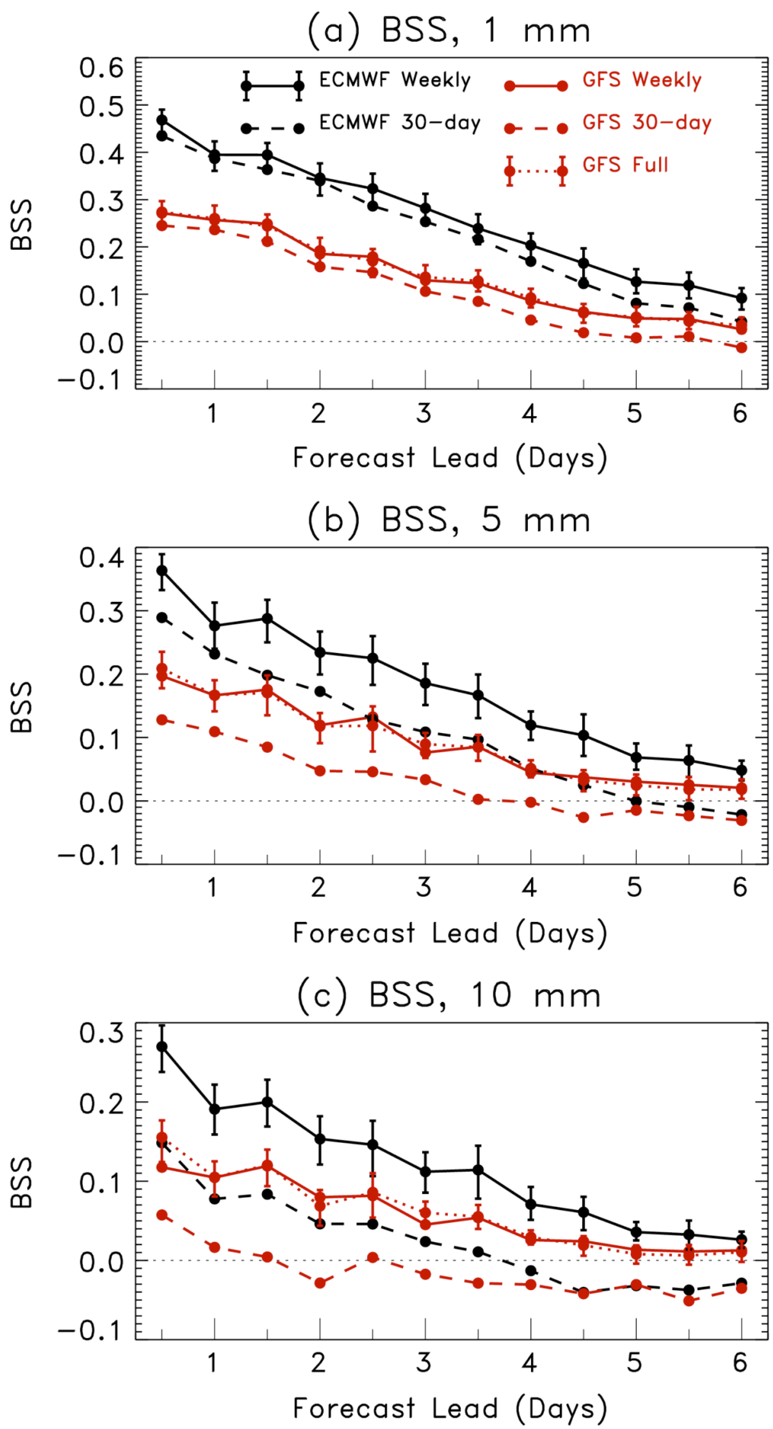


Precipitation skill with weekly, 30-day, and full training data sets

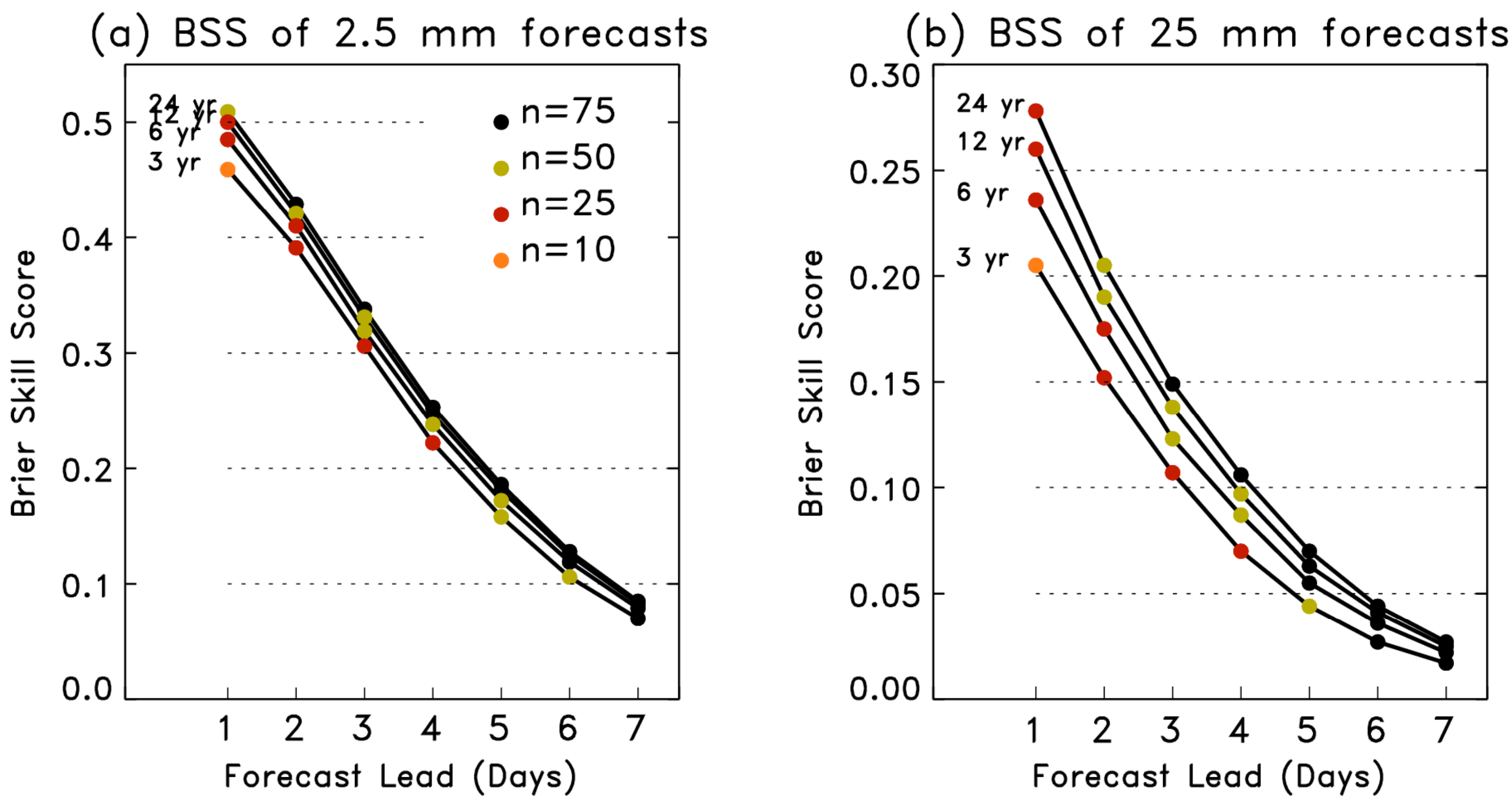
Notes:

(1) Substantial benefit of weekly relative to 30-day training data sets, especially at high thresholds.

(2) Not much benefit from full relative to weekly reforecasts.



Effect of training sample size: previous results with GFS



colors of dots indicate which size analog ensemble provided the largest amount of skill.

Conclusions

- Reforecasts shown to aid in calibration of forecasts for a wide variety of applications
- Still a large benefit from forecast calibration, even with state-of-the-art ECMWF forecast model.
  - Temperature calibration:
    - Short leads: a few previous forecasts adequate for calibration
    - Long leads: better skill with long reforecast training data set.
  - Precipitation calibration
    - Low thresholds: a few previous forecasts somewhat ok for calibration
    - Larger thresholds: large benefit from large training data set.
    - Skill when trained with daily data not much larger than when trained with weekly data (preliminary result, more testing needed).

Are operational centers heading toward reforecasting?

- **NCEP**: tentative plans for 1-member real-time reforecast.
- **ECMWF**: once-weekly, real-time 5-member reforecast starting mid 2008.
- **RPN Canada**: planning ~5-year reforecast data set, delayed by budget and staffing issues.

References

- Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2007: Probabilistic forecast calibration using ECMWF and GFS ensemble forecasts. Part I: surface temperature. *Mon. Wea. Rev*., in press. Available at http://tinyurl.com/3axuac
- Hamill, T. M., J. S. Whitaker, and R. Hagedorn, 2007: Probabilistic forecast calibration using ECMWF and GFS ensemble forecasts. Part II: precipitation. *Mon. Wea. Rev*., in press. Available at http://tinyurl.com/38jgkv
- (and many reforecast references therein)